Lineaire regressie
Variantie- en covariantie-analyse

Methoden in de psychologie

Academiejaar 2020-2021

Inhoudsopgave

1 Het lineair regressiemodel

1.1 Lineaire regressie

Regressie is een statistische techniek om het verband tussen één (*univariaat*) of meerdere (*multivariaat*) **uitkomst(en)** en een set van **predictoren** te onderzoeken. Een regressiemodel is een hypothetisch statistisch model dat de relatie tussen de uitkomst en de predictoren beschrijft. De invloed van de predictoren op de uitkomst(en) wordt gemodelleerd.

In dit hoofdstuk beschouwen we **univariate regressie**.

De uitkomst is de te verklaren variabele of **afhankelijke** variabele (Y) en de predictoren zijn de **onafhankelijke** variabelen (X). Men spreekt van regressie van Y op X. Y is gemeten op minstens intervalniveau.

We werken met het softwarepakket R. Alle data en R-code voor de voorbeelden uit de cursus worden ter beschikking gesteld (zie verder).

Lineaire regressie kan gebruikt worden om:

- het effect van predictoren op de uitkomst of de samenhang tussen uitkomst en predictoren na te gaan.
- toekomstige observaties te voorspellen, gegeven de waarden voor de predictoren.
- een algemene beschrijving van de datastructuur weer te geven.

Lineaire regressie is al aan bod gekomen in Statistiek II. We bouwen er hier op verder. Sommige zaken worden herhaald om nadien verder uit te breiden. Niet alles wat in Statistiek II aan bod komt, wordt hier expliciet herhaald, maar wordt wel beschouwd als voorkennis. Er zullen bvb. zaken aan bod komen in de oefenlessen die niet meer aan bod komen in de theorie, maar die geziene stof uit Statistiek II zijn.

Een lineair regressiemodel met p predictoren voor n onafhankelijke observaties wordt als volgt voorgesteld:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n \\ Y_i &= \beta_0 + \sum_{\ell=1}^{p} \beta_\ell x_{i\ell} + \varepsilon_i \qquad i = 1, \ldots, n \end{aligned} \tag{1}$$

- $x_{i\ell}$: score voor de predictor X_ℓ voor observatie i ($\ell = 1, \ldots, p$)

- β_0, β_1, ..., β_p: regressiecoëfficiënten (populatieparameters)

Het stochastisch deel van model (1) bestaat uit 3 assumpties met betrekking tot de fouttermen:

1. $E(\varepsilon_i) = 0$ voor alle i
2. $\text{Var}(\varepsilon_i) = \sigma^2$ voor alle i, i.e. constante variantie van de fouttermen of *homoscedasticiteit*
3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ voor alle $i \neq j$

Bijkomende veronderstellingen: de predictoren zijn onafhankelijk van ε_i, de predictoren zijn zonder fout gemeten.

Onbekende parameters: β_0, β_1, ..., β_p, σ^2. In totaal zijn dit $p+2$ parameters, $p+1$ regressiecoëfficiënten en σ^2.

$E(Y_i|x_{i1}, x_{i2}, \ldots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$
$\Rightarrow$ Betekenis β_ℓ ($\ell = 1, \ldots, p$): als de ℓ-de predictor (X_ℓ) met 1 eenheid stijgt terwijl alle overige predictoren constant blijven dan neemt de verwachte waarde van Y toe met β_ℓ.

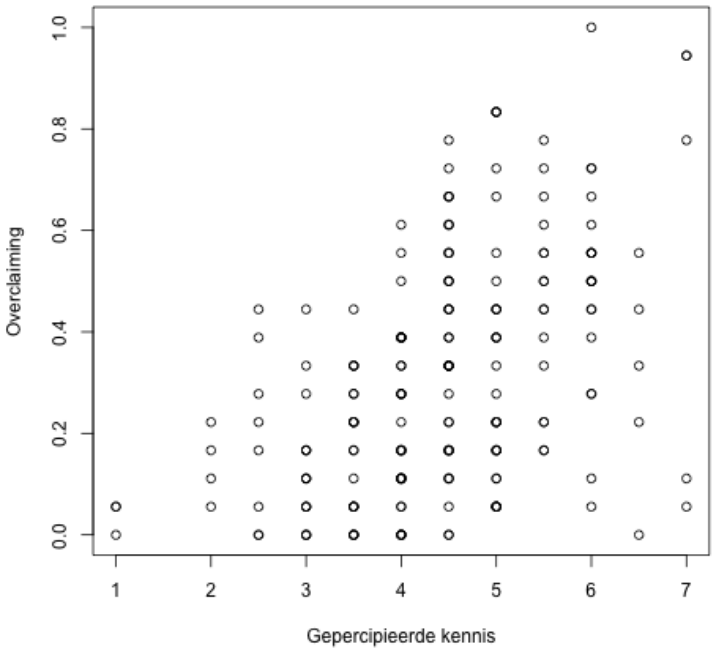
De term 'lineair' krijgt vaak een dubbele betekenis:

1. Primaire betekenis: het regressiemodel is lineair in de parameters.
   Model (1) is lineair in de parameters. De parameters komen niet voor in een niet-lineaire vorm zoals β_1^2, $\log(\beta_1)$, $\exp(\beta_2), \ldots$.
2. Vaak gebruikt men de term echter ook om naar de aard van de samenhang tussen de variabelen te verwijzen. In veel toepassingen gebruikt men immers een lineaire regressiefunctie, dit betekent dat het model ook lineair is in de predictor(en) X. Dit is niet noodzakelijk.
   $Y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ ($i = 1, \ldots, n$) is, in tegenstelling tot model (1), een statistisch model met een niet-lineair verband tussen de uitkomst en predictor, het verband is kwadratisch. Toch is dit model een lineair regressiemodel aangezien het lineair is in de parameters.

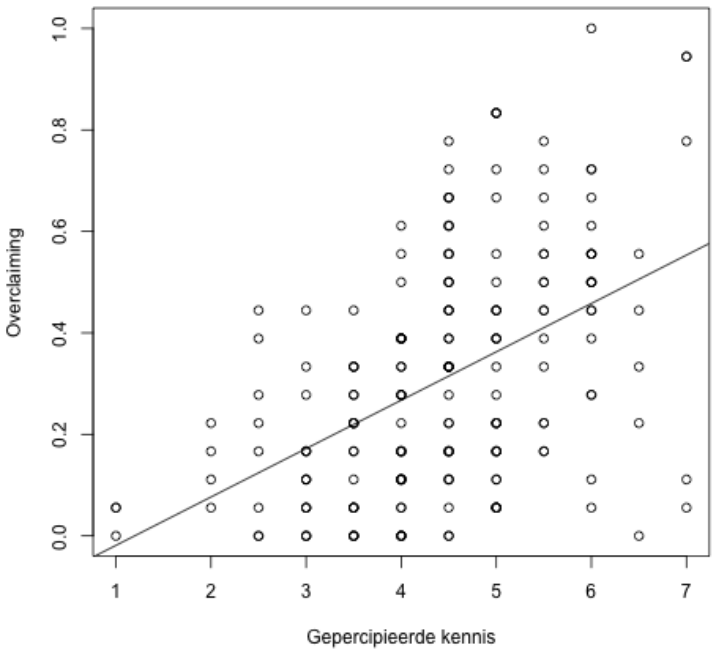
De term regressie impliceert vaak dat ook alle onafhankelijke variabelen van intervalniveau zijn. Dit is echter niet strikt noodzakelijk. We beschouwen ook regressie met enkel nominale onafhankelijke variabelen en regressie met onafhankelijke variabelen van zowel nominaal als intervalniveau.

Bij de start van de data-analyse is het nuttig om een *scatterplot* of *spreidingsdiagram* te maken. Hierbij wordt de afhankelijke variabele Y op de y-as gezet en de onafhankelijke

variabele X op de x-as. Veronderstel dat men geïnteresseerd is in de mate van overclaiming in functie van gepercipieerde kennis (zie sectie 1.2.1 voor meer uitleg over de data).



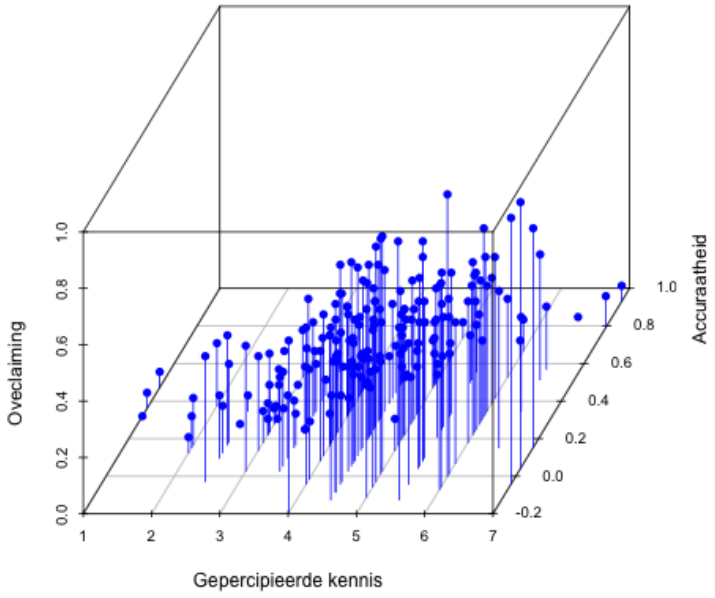
Met behulp van lineaire regressie kunnen we de best passende rechte door de puntenwolk bepalen, d.i. de rechte die zo goed mogelijk de trend van de gegevens benadert.



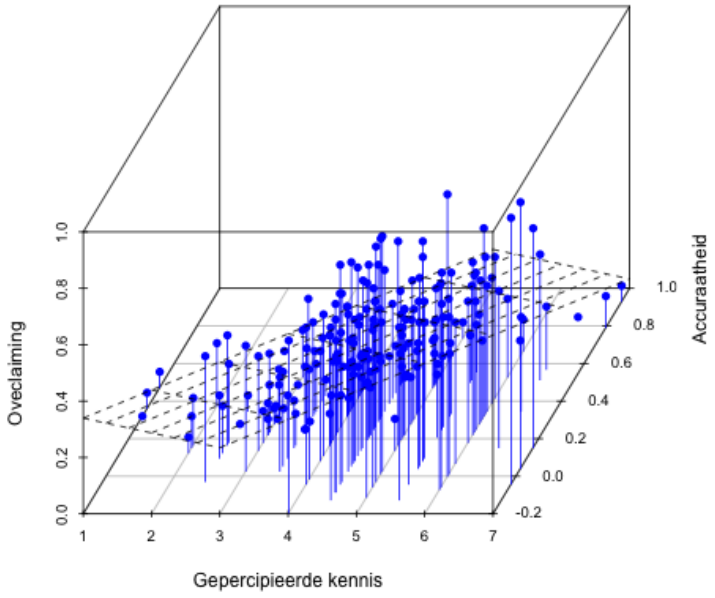
Algemeen kan elk lineair regressieprobleem waarbij p onafhankelijke variabelen of predictoren beschouwd worden, grafisch voorgesteld worden in $p + 1$ dimensies.

Wanneer we 2 predictoren beschouwen kunnen we analoog een 3-dimensioneel spreidingsdiagram maken en daardoor het best passende vlak tekenen.

Veronderstel dat we in bovenstaand voorbeeld de mate van overclaiming voorspellen in functie van gepercipieerde kennis en accuraatheid, dan ziet het 3-dimensioneel spreidingsdiagram er als volgt uit:



Met behulp van lineaire regressie kunnen we het best passende vlak door de puntenwolk bepalen:



In deze cursusnota's gaan we aan de slag met enkele voorbeelden die we gebruiken om een aantal technische aspecten m.b.t. lineaire regressie te bespreken. Niet alle zaken die aan bod kwamen in Statistiek I en Statistiek II worden herhaald. Zo zijn datavisualisaties, het univariaat verkennen van de dataset, het bekijken van onderlinge relaties tussen variabelen essentieel, alsook het nagaan van de assumpties die aan de basis liggen van de analyses. Aan deze zaken wordt de nodige aandacht besteed in de oefenlessen.

In de volgende sectie worden de voorbeelden die we doorheen de cursus gaan gebruiken, geïntroduceerd.

1.2 Voorbeelden

1.2.1 Overclaiming

We beschouwen hier data uit een studie uit de volgende paper:

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds:

Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science, 26*, 1295-1303.

Bron: Open Stats Lab
https://sites.trinity.edu/osl/data-sets-and-activities/regression-activities

Zowel de code die hoort bij de analyses (`overclaiming.R`) als de data (`overclaiming.csv`) zijn terug te vinden op Ufora.

Achtergrond

Mensen kunnen hun eigen kennis overschatten, soms zelfs kennis van concepten, gebeurtenissen en mensen die niet bestaan en dus niet gekend kunnen zijn. Dit fenomeen heet *overclaiming*.

In deze studie wenst men na te gaan in welke mate zelf-gepercipieerde kennis de mate van overclaiming voorspelt.

Methode en design

202 participanten nemen deel aan de studie (85 vrouwen, 115 mannen, 2 personen van wie gender niet gekend is; gemiddelde leeftijd is 33.5 jaar met een standaarddeviatie gelijk aan 10.0). De steekproef werd getrokken uit een lijst van Amazon en het betreft enkel particpanten uit de Verenigde Staten.

De participanten vullen een vragenlijst in om hun algemene kennis rond persoonlijke financiën te scoren en een test om overclaiming te bepalen. Beide testen worden afgenomen in een countergebalanceerde volgorde over participanten.

Bij de overclaimingtaak worden 15 items gepresenteerd in een random volgorde, waarvan 12 items gaan over bestaande termen en 3 items over niet-bestaande termen uitgevonden door de onderzoekers.
We are interested in common knowledge about personal finance. You will see 15 terms related to personal finance. Please rate your knowledge about each term by choosing the appropriate number from 1 (never heard of it) to 7 (very knowledgeable).

Daarna wordt bij de participanten ook nog een test afgenomen rond financiële geletterdheid (FINRA Investor Education Foundation).

Data

De dataset bevat de volgende variabelen:

order_of_tasks De volgorde van de taken: `order_of_tasks=1` wanneer de participanten eerst de test rond gepercipieerde kennis afleggen en dan de test voor overclaiming; `order_of_tasks=2` wanneer de participanten eerst de test voor overclaiming afleggen en dan de test voor gepercipieerde zelfkennis

In de analyses brengen wij deze variabele niet mee in rekening.

self_perceived_knowledge Score voor gepercipieerde kennis, wordt als van intervalniveau verondersteld

De vragen voor algemene kennis rond persoonlijke financiën zijn als volgt:

> In general, how knowledgeable would you say you are about personal finance? (1 = not knowledgeable at all, 7 = extremely knowledgeable)
>
> How would you rate your general knowledge of personal finance compared to the average American? (1 = much less knowledgeable, 7 = much more knowledgeable)

Om de score te bekomen, wordt het gemiddelde genomen van het antwoord op beide vragen.

overclaiming_proportion Score voor overclaiming, wordt als van intervalniveau verondersteld

Overclaiming wordt gemeten door na te gaan voor welke proportie van onbestaande termen een participant kennis beweert te hebben. Deze proportie wordt berekend voor 6 cutoffs voor kennis: de proportie van onbestaande termen met een score van 2 of hoger, de proportie van onbestaande termen met een score van 3 of hoger enz. voor 4, 5, 6 en 7. Daarna wordt het gemiddelde genomen van deze proporties.

accuracy Accuraatheid, wordt als van intervalniveau verondersteld

Accuraatheid wordt bekomen door analoog als bij het bepalen van overclaiming, de proportie te bepalen van bestaande termen waarover de participant kennis beweerde te hebben en hiervan de proportie van onbestaande termen af te trekken.

FINRA_score Score op de test voor financiële geletterdheid, wordt als van intervalniveau verondersteld (indicator voor werkelijke kennis)

Onderzoeksvraag

De onderzoekers wensen na te gaan in welke mate gepercipieerde kennis overclaiming voorspelt, rekening houdend met (i.e. controlerend voor) accuraatheid en FINRA-score.

Analyses met deze data in deze cursusnota's kunnen teruggevonden worden op pagina 18, 23, 26, 29, 32, 37, 39, 41, 52, 54 en 56.

1.2.2 Pijneducatie bij onderrugpijn

We beschouwen hier een voorbeeld van een quasi-experiment. De fictieve data zijn gebaseerd op de volgende studie:

Mosely, G.L., 2004, Evidence for a direct relationship between cognitive and psychological change during an education intervention in people with chronic low back pain. *European Journal of Pain, 8*, 39-45

In de oefeningensessies wordt hetzelfde voorbeeld hernomen.

Zowel de code die hoort bij de analyses (`pijneducatie.R`) als de data (`pijneducatie.csv`) zijn terug te vinden op Ufora.

Achtergrond

Men stelt vast dat naast puur fysieke factoren ook cognitieve factoren een rol spelen bij pijnperceptie. In deze studie stelt men zich de vraag of deze cognities een *actieve* rol spelen in de pijnperceptie bij mensen met klachten over pijn in de onderrug.

Bovendien bestaat er in de literatuur ook evidentie voor verschillende soorten van betrokken cognities: enerzijds meer algemene cognities over pijn en anderzijds meer specifieke pijn-locatie gerelateerde cognities.

Om dit in meer detail na te gaan voeren de onderzoekers een manipulatie van verschillende pijn-gerelateerde cognities uit en gaan ze na wat de effecten op pijnperceptie zijn. Daarnaast wensen de auteurs ook te controleren voor comorbiditeit van depressie en leeftijd.

Men wenst te onderzoeken of er evidentie is voor het bestaan van verschillen tussen de onderliggende cognities bij pijnperceptie.

In een poging om een zuivere pijnindicator te definiëren, wordt bij de patiënten gemeten hoe ver ze voorover kunnen buigen.

Methode en design

De patiënten zijn geselecteerd via een lokaal ziekenhuis ($n = 121$) waar één van de onderzoekers makkelijk toegang tot heeft. Subjecten worden geweerd indien er comorbiditeit met andere fysieke of zuiver neurologische aandoeningen is. Alle patiënten ondergaan een één-op-één sessie met een therapeut waarin een uitgebreide pijneducatie sessie plaats vindt.

Condities (pijn-educatie groepen)

We onderscheiden de volgende condities:

- Conditie 1: hier ligt de focus op cognities die gerelateerd zijn aan de algemene pijnfysiologie.
  Deze conditie belicht de werking van pijn-, druk- en andere receptoren binnen het centrale en perifere zenuwstelsel.
- Conditie 2: de nadruk ligt hier op de fysiologie van de ruggengraat.
  Er wordt belicht hoe de verschillende wervels samenzitten in de ruggengraat.
- Conditie 3: dit is de baseline conditie.
  In deze conditie gaat men dieper in op de algemene werking van het maag- en darmstelsel.

Toewijzing

De eerste 41 patiënten worden toegewezen aan de eerste conditie, de volgende 40 aan de tweede conditie. Tot slot worden de laatste 40 patiënten toegewezen aan de baseline conditie.

Data

Afhankelijke variabele:

> Er wordt bij iedere deelnemer 2 keer gemeten hoe ver men voorover kan buigen (in mm): 1 keer vóór en 1 keer na de experimentele conditie (conditie 1, 2 of 3). Er wordt vervolgens een verschilscore berekend (na-vóór). Positieve verschilscores wijzen op een verbetering.

Onafhankelijke variabelen en predictoren:

- Een nominale variabele duidt de 3 condities aan.
- Leeftijd van de deelnemers
- Depressiescore; deze score mag als van intervalniveau beschouwd worden.

De dataset bevat bijgevolg de volgende variabelen:

Buig Verschilscore hoe ver iemand voorover kan buigen (na-vóór), in mm uitgedrukt

Gender Geslacht; Vrouw =1, man =0

Leeft Leeftijd in jaren

Conditie Conditie=1: algemene pijn-fysiologie groep, Conditie=2: Ruggengraat-educatie groep, Conditie=3: baseline groep

Dep Depressiescore

Onderzoeksvragen

Men wil nagaan of er een verschil is in gemiddelde uitkomst tussen de verschillende onderliggende cognities bij pijnperceptie.

Aangezien het hier geen gerandomiseerde studie betreft (deelnemers zijn niet at random toegewezen aan de condities) is het zinvol om het onderzochte effect ook te corrigeren voor het effect van leeftijd en de graad van depressie. Er kan ook nagegaan worden of effecten variëren volgens leeftijd en/of depressie.

Analyses met deze data in deze cursusnota's kunnen teruggevonden worden op pagina 44, 57 en 59.

1.2.3 Herstel na coma

We gebruiken een subset van de data uit de volgende paper:

Wong, P. P., Monette, G., & Weiner, N. I. (2001) Mathematical models of cognitive recovery. *Brain Injury, 15*, 519530.

Zowel de code die hoort bij de analyses (`coma.R`) als de data (`coma.csv`) zijn terug te vinden op Ufora.

Achtergrond

De data die we gebruiken zijn een onderdeel van een longitudinale studie waarin modellen voor IQ opgesteld worden om herstelverloop na een coma te voorspellen.

200 patiënten die als gevolg van een traumatisch hersenletsel in coma lagen voor bepaalde tijd, worden na het ontwaken opgevolgd en er worden periodisch standaard IQ-testen afgenomen. Deze testen kunnen afgenomen worden na de coma, na het uiteindelijke herstel en op alle tijdspunten tussenin. Op die manier kan het herstelverloop van IQ onderzocht worden.

Wij beschouwen voor iedere patiënt slechts 1 meting.

De originele volledige dataset is beschikbaar in het R-package `carData`.

Data

Wij beschouwen de volgende variabelen:

duration Duur van de coma (in dagen)

sex Geslacht van de patiënt

age Leeftijd (in jaren) op moment van trauma

piq Mathematisch (*performance*) IQ

viq Verbaal IQ

duration_cat Nominale variabele met categorieën volgens duur van coma

> Categorieën zijn: t.e.m. 1 dag, meer dan 1 dag t.e.m. 7 dagen, meer dan 7 dagen t.e.m. 14 dagen, meer dan 14 dagen t.e.m. 255 dagen

Wij gebruiken de dataset hoofdzakelijk op een exploratieve manier om interacties tussen variabelen beter te begrijpen. We beschouwen hierbij een lineair regressiemodel met mathematisch IQ (`piq`) als afhankelijke variabele.

Analyses met deze data in deze cursusnota's kunnen teruggevonden worden in sectie 5.4 op pagina 70.

1.2.4 Levenstevredenheid

We beschouwen hier een voorbeeld met fictieve data uit de paper van Preacher & Hayes (2004):

Preacher, K.J., Hayes, & A.F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*, 717-731.

Men is geïnteresseerd in het onderzoeken van het proces of de manier waarop cognitieve gedragstherapie een invloed heeft op levenstevredenheid na pensionering bij klinisch depressieve personen.

Zowel de code die hoort bij de analyses (`satisfaction.R`) als de data (`satisfaction.csv`) zijn terug te vinden op Ufora.

Methode en design

Dertig klinisch depressieve patiënten worden at random toegewezen aan de conditie die bestaat uit 10 sessies van de nieuwe therapie of aan de conditie die bestaat uit 10 sessies van een standaardtherapie .

Na sessie 8 wordt de positiviteit bepaald van de attributies die de deelnemers maken voor een recente negatieve ervaring (attribueren = toekennen van oorzaken). Op het einde van sessie 10 wordt de algemene levenstevredenheid gemeten.

Data

De dataset bevat de volgende variabelen:

satis Algemene levenstevredenheid (wordt als van intervalniveau verondersteld), gemeten op sessie 10

therapy Therapie; `therapy = 1`: nieuwe therapie, `therapy = 0`: standaardtherapie

attrib Attributie (wordt als van intervalniveau verondersteld), gemeten op sessie 8

Onderzoeksvraag

De onderzoeksvraag is of het eventuele effect van de cognitieve therapie op de algemene levenstevredenheid (deels) verloopt via attributie, m.a.w. of cognitieve therapie een invloed heeft op attributie die dan op zijn beurt een invloed heeft op de algemene levenstevredenheid. De onderzoeksvraag is dus of het effect van de cognitieve gedragstherapie *gemedieerd* wordt door de positiviteit van de attributies.

Analyses met deze data in deze cursusnota's kunnen teruggevonden worden op pagina 93.

1.3 Matrixnotaties

Aangezien een lineair regressiemodel met p predictoren voor n observaties als volgt is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n$$

houdt dit in:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{np} + \varepsilon_n \end{aligned}$$

Aan de hand van matrixnotaties kunnen we dit model compact als volgt noteren:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

waarbij:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

We hebben dus

$$\underbrace{\mathbf{Y}}_{n\times 1} = \underbrace{\mathbf{X}}_{n\times(p+1)} \underbrace{\boldsymbol{\beta}}_{(p+1)\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{n\times 1}$$

De matrix $\mathbf{X}$ wordt de **designmatrix** genoemd. Deze matrix is gedefinieerd als een kolom van 1-en en p kolommen die telkens de n observaties van de predictoren $X_1, X_2, \ldots, X_p$ bevatten. De kolom met 1-en wordt gebruikt als het lineaire model een intercept bevat (d.i. als $\beta_0 \neq 0$).

Wanneer we in het voorbeeld rond overclaiming (zie sectie 1.2.1), de uitkomst overclaiming regresseren op gepercipieerde kennis, accuraatheid en FINRA-score, dan ziet de designmatrix er als volgt uit (we tonen hier enkel de vier eerste rijen):

```
  (Intercept) self_perceived_knowledge    accuracy FINRA_score
1           1                      5.5  0.25000000           4
2           1                      4.5  0.19444444           4
3           1                      3.5  0.34722222           5
4           1                      6.0 -0.05555556           4
```

Verder geldt dat $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ aangezien:

$$E(\boldsymbol{\varepsilon}) = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \vdots \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Hieruit volgt dat $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$.

Aangezien $\text{Var}(\varepsilon_i) = \sigma^2$ voor alle i en $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ voor alle $i \neq j$, hebben we:

$$\text{Var}(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 \mathbf{I}.$$

1.4 Parameterschattingen

β_0, β_1, …, β_p en σ^2 zijn onbekende parameters.

Op basis van n observaties wensen we de onbekende populatieparameters te schatten.

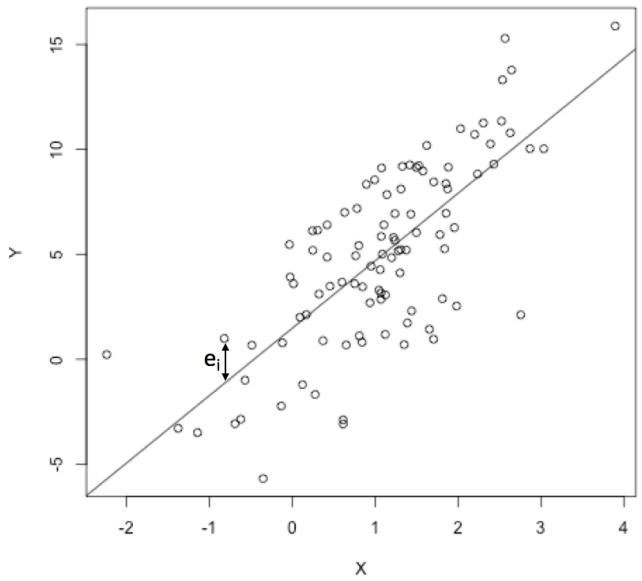
Concreet: gegeven een lukraak getrokken steekproef van n observaties $Y_1, \dots, Y_n$ en bijhorende predictoren $x_{1\ell}, \dots, x_{n\ell}$ ($\ell = 1, \dots, p$) wensen we β_0, β_1, …, β_p en σ^2 te schatten.

Twee courante methodes:

- Methode van de kleinste kwadraten
- Methode van de maximale aannemelijkheid

1.4.1 Kleinste kwadratenschatters voor β

Grafisch (2 dimensies, d.i. enkelvoudig):



De best passende rechte wordt door de puntenwolk getekend.

De kleinste kwadraten regressielijn is een unieke rechte die men bekomt door de som van de verticale kwadratische afstanden (afwijkingen) tussen ieder datapunt en de rechte te minimaliseren.

Analoog voor meerdere dimensies: de regressievergelijking die men bekomt via de methode van de kleinste kwadraten is de vergelijking van het hypervlak waarvoor de som van de kwadratische afwijkingen tussen ieder datapunt en dat vlak geminimaliseerd is.

Met **B** duiden we de vector van de kleinste kwadratenschatters aan:

$$\begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_p \end{bmatrix}.$$

B_0, B_1, …, B_p zijn de **schatters** voor respectievelijk β_0, β_1, …, β_p.

Schatters zijn kansvariabelen en worden dus genoteerd met grote letters: wanneer we herhaaldelijk een andere lukrake steekproef nemen waarbij de predictoren constant gehouden worden, zouden we telkens andere waarden bekomen voor deze schatters. De verdeling van schatters noemt men de **steekproevenverdeling**.

b_0, b_1, ..., b_p zijn de **puntschattingen**, i.e. een concreet getal in een gegeven steekproef.

Noteer $\hat{Y}_i = B_0 + B_1 x_{i1} + B_2 x_{i2} + \dots B_p x_{ip}$, $\hat{Y}_i$ is de gefitte waarde van Y_i. Merk op dat $\hat{Y}_i$ een schatter is voor $E(Y_i|x_{i1}, \dots, x_{ip})$ en niet voor Y_i!

De vector van gefitte waarden wordt voorgesteld door $\hat{\mathbf{Y}}$:

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{XB}.$$

$E_i = Y_i - \hat{Y}_i$ is de afwijking tussen Y_i en $\hat{Y}_i$, E_i is het residu van Y_i.

Technisch gezien minimaliseert de kleinste kwadratenmethode $\sum_{i=1}^{n} E_i^2$.

$\sum_{i=1}^{n} E_i^2$ is de **residuele** of **fout kwadratensom**, in het Engels wordt deze term *residual/error sum of squares* genoemd of kortweg SSE [1].

De SSE kan men beschouwen als een maat voor de fout van het regressiemodel, of nog, een maat voor het verlies aan informatie door Y_i te vervangen door $\hat{Y}_i$.

De vector van residuen wordt voorgesteld door $\mathbf{E}$

$$\mathbf{E} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{bmatrix} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

Er kan aangetoond worden dat:

$$\underbrace{\mathbf{B}}_{(p+1)\times 1} = \underbrace{(\mathbf{X}'\mathbf{X})^{-1}}_{(p+1)\times(p+1)} \underbrace{\mathbf{X}'\mathbf{Y}}_{(p+1)\times 1}$$

Er bestaat slechts een unieke oplossing voor de kleinste kwadratenschatters indien de inverse van de matrix $\mathbf{X}'\mathbf{X}$ bestaat. Dit betekent dat deze matrix van volledige rang of niet-singulier moet zijn.

[1] In de literatuur wordt zowel de term 'residual sum of squares' als 'error sum of squares' gebruikt. Beiden zijn gelijk. Er geldt dus SSE=SS_{Res}.

Waarom zijn kleinste kwadratenschatters goede schatters?

1. Ee schatters hebben een zinvolle meetkundige betekenis (denk aan het minimaliseren van de som van de kwadratische afwijkingen van de datapunten tot de regressierechte in 2 dimensies).
2. Wanneer de fouttermen (ε_i) onafhankelijk en identisch normaal verdeeld zijn, zijn de kleinste kwadratenschatters gelijk aan de **maximum likelihood schatters**. Dit zijn de schatters die bekomen worden via de methode van de maximale aannemelijkheid.

   Via maximum likelihood worden de waarden voor de parameters gekozen die het meest aannemelijk zijn, gegeven de geobserveerde data.
3. Het Gauss-Markov theorema stelt dat de kleinste kwadratenschatters de beste lineaire zuivere schatters zijn (zie volgende sectie).

1.4.2 Het Gauss-Markov theorema

Het Gauss-Markov theorema stelt:

Indien $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ en de structurele component van het model correct is ($E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$) dan zijn de kleinste kwadratenschatters **zuiver**, **efficiënt** en **lineair**.

- Het **zuiver** of **onvertekend** zijn van de schatters impliceert:

  $$E(\mathbf{B}) = \boldsymbol{\beta}.$$

  Concreet: $E[B_0] = \beta_0$, $E[B_1] = \beta_1$, …, $E[B_p] = \beta_p$.
  Dit betekent dat de verwachte waarde van de schatters gelijk is aan de regressiecoëfficiënten (populatieparameters), de gezochte parameters.

- De schatters zijn **efficiënt** (of nog: de 'beste') omdat hun variantie (i.e. de variantie van hun steekproevenverdeling) minimaal is (in vergelijking met alle *onvertekende lineaire* schatters).

  $$\text{Var}(\mathbf{B}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

  σ^2 is ongekend en moet ook geschat worden (zie volgende sectie). Merk opnieuw op dat de standaarddeviatie (vierkantswortel van de variantie) van een schatter ook de **standaardfout** genoemd wordt.

- De schatters noemt men **lineair** omdat ze een lineaire functie zijn van de observaties Y_i.

Het theorema toont aan dat de kleinste kwadratenschatters een goede keuze zijn, maar wanneer de fouttermen bvb. gecorreleerd zijn of ongelijke varianties hebben, bestaan er betere schatters. We gaan daar in deze cursus niet verder op in.

1.4.3 Een schatter voor σ^2

Een schatter voor σ^2 kan als volgt bekomen worden:

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-(p+1)} = \frac{\sum_{i=1}^{n} E_i^2}{n-(p+1)} = \frac{\text{SSE}}{n-(p+1)}.$$

In totaal worden $p+1$ regressiecoëfficiënten geschat, namelijk β_0, β_1, …, β_p. Dit betekent dat dat we $p+1$ vrijheidsgraden verliezen, daarom delen we door $n-(p+1)$.

S^2 wordt ook wel MSE genoemd (*Mean Squared Error*[2]).

S^2 is een onvertekende schatter voor σ^2: $E[S^2] = \sigma^2$.

1.4.4 Voorbeeld: overclaiming

We bekijken het effect van gepercipieerde kennis op overclaiming (zie sectie 1.2.1) op basis van een lineair regressiemodel.

```
> fit1_expertise<-lm(overclaiming_proportion~self_perceived_knowledge,data=expertise)
> fit1_expertise

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge,
data = expertise)

Coefficients:
(Intercept)  self_perceived_knowledge
-0.11406                     0.09532
```

We lezen af dat $b_0 = -0.114$ en $b_1 = 0.0953$. Omdat b_1 positief is, hebben we een stijgende geschatte regressierechte: er is een (al dan niet significante) stijgende trend of een positief verband tussen gepercipieerde kennis en overclaiming. We schatten dat, indien gepercipieerde

[2] Alternatieve notatie: MS_{Res}

kennis met 1 eenheid stijgt, de gemiddelde proportie overclaiming met 0.0953 eenheden toeneemt.

Via de functie `summary` krijgen we meer informatie over het regressiemodel.

```
> summary(fit1_expertise)

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge,
data = expertise)

Residuals:
Min       1Q   Median       3Q      Max
-0.50551 -0.15610  0.00662  0.12167  0.54215

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.11406    0.05624  -2.028   0.0439 *
self_perceived_knowledge  0.09532    0.01228   7.762 4.22e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2041 on 200 degrees of freedom
Multiple R-squared:  0.2315,Adjusted R-squared:  0.2277
F-statistic: 60.25 on 1 and 200 DF,  p-value: 4.225e-13
```

In de kolom `Estimate` staan de geschatte regressiecoëfficiënten. In de kolom `Std. Error` staan de geschatte standaardfouten. Hier is $s_{B_0} = 0.0562$ en $s_{B_1} = 0.0123$. We lezen verder af dat de `Residual standard error` gelijk is aan 0.204, dit is s met s^2 een schatting voor σ^2.

Hoewel de onderzoekers geïnteresseerd zijn in het verband tussen overclaiming en gepercipieerde kennis, zijn er meerdere mogelijke predictoren voor overclaiming. We voegen enkele van deze predictoren toe aan het regressiemodel. Dit laat o.a. toe om het effect van gepercipieerde kennis te bekijken, conditioneel op de andere predictoren.

```
> fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
                        +FINRA_score,data=expertise)
> summary(fit3_expertise)

Call:
```

```
lm(formula = overclaiming_proportion ~ self_perceived_knowledge +
accuracy + FINRA_score, data = expertise)

Residuals:
Min       1Q   Median       3Q      Max
-0.38033 -0.08672 -0.01418  0.08808  0.30886

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.057787   0.039203   1.474   0.1421
self_perceived_knowledge  0.094069   0.008018  11.732   <2e-16 ***
accuracy                 -0.793219   0.045655 -17.374   <2e-16 ***
FINRA_score               0.018370   0.008576   2.142   0.0334 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1256 on 198 degrees of freedom
Multiple R-squared:  0.7116,Adjusted R-squared:  0.7073
F-statistic: 162.9 on 3 and 198 DF,  p-value: < 2.2e-16
```

We zien nog steeds een positief verband tussen gepercipieerde kennis en overclaiming: we schatten dat voor een constante accuraatheid en FINRA-score, de gemiddelde proportie overclaiming met 0.0940 eenheden toeneemt als gepercipieerde kennis met 1 eenheid stijgt.

We zien verder een negatief effect van accuraatheid: we schatten dat voor een constante gepercipieerde kennis en FINRA-score, de gemiddelde proportie overclaiming met 0.793 eenheden afneemt als accuraatheid met 1 eenheid stijgt. Hoewel deze interpretatie wiskundig correct is, is ze niet zo nuttig in dit geval. De verdeling van de variabele `accuracy` ziet er als volgt uit:

```
> summary(expertise$accuracy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.1944  0.1562  0.2847  0.2953  0.4549  0.9306
```

Een toename van 1 eenheid is dus niet betekenisvol. Het is interessanter om het effect te bekijken van bvb. een toename van 0.10 eenheden in accuracy (terwijl de overige predictoren constant blijven). We schatten in dat geval dat de gemiddelde proportie overclaiming met $0.793 \times 0.10 = 0.0793$ eenheden afneemt.

1.5 Predictie

1.5.1 Predicties

Gegeven de schatters voor de regressiecoëfficiënten, kunnen we op basis van het regressiemodel de (verwachte) waarde van Y_* voorspellen op basis van nieuwe (niet eerder geobserveerde) waarden voor de set van predictoren x_{1*}, x_{2*}, ..., x_{p*}:

$$\hat{Y}_* = B_0 + B_1 x_{1*} + B_2 x_{2*} + \ldots + B_p x_{p*}.$$

"*Wat is de verwachte mate van overclaiming, gegeven de scores voor gepercipieerde kennis en accuraatheid en de FINRA-score?*"

Vaak wenst men predicties te doen voor een 'typisch' profiel. Dan kan men bvb. een predictie doen waarbij elke predictor gelijk gesteld wordt aan het steekproefgemiddelde. We hernemen het voorbeeld rond overclaiming (zie sectie 1.2.1).

```
> mean(expertise$self_perceived_knowledge)
[1] 4.428218
> mean(expertise$accuracy)
[1] 0.2953108
> mean(expertise$FINRA_score)
[1] 3.69802
```

De geschatte regressiecoëfficiënten van het model waarbij overclaiming geregresseerd wordt op gepercipieerde kennis, accuraatheid en FINRA-score, zijn als volgt:

```
> fit3_expertise

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge +
accuracy + FINRA_score, data = expertise)

Coefficients:
(Intercept)  self_perceived_knowledge                 accuracy
0.05779                   0.09407                  -0.79322
FINRA_score
0.01837
```

De voorspelde mate van overclaiming voor een typisch profiel is bijgevolg gelijk aan $0.05779 + 0.09407 \times 4.428218 - 0.79322 \times 0.2953108 + 0.01837 \times 3.69802 = 0.308$.

In R kunnen we predicties op een eenvoudige manier bekomen door een dataframe te maken waarin de waarden voor de predictoren ingevuld worden en op basis van het geschatte model de uitkomst te voorspellen voor de waarden in de nieuwe dataframe.

```
>avprofiel<-data.frame(self_perceived_knowledge=mean(expertise$self_perceived_knowledge),
                       accuracy=mean(expertise$accuracy),FINRA_score=mean(expertise$FINRA_score))
> predict(fit3_expertise,newdata=avprofiel)
1
0.3080308
```

Bij het voorspellen van de uitkomst op basis van nieuwe waarden voor de set van predictoren x_{1*}, x_{2*}, ..., x_{p*} in het model, kunnen we een onderscheid maken tussen 2 zaken:

- het voorspellen van de verwachting van een toekomstige waarde : $E(Y_*|x_{1*}, x_{2*}, \ldots, x_{p*})$, een gemiddelde
- het voorspellen van een toekomstige waarde: Y_*, een individuele observatie

De schatters voor beide gevallen zijn dezelfde en gelijk aan
$\hat{Y}_* = B_0 + B_1 x_{1*} + B_2 x_{2*} + \ldots + B_p x_{p*}$.

Bij intervalschattingen kunnen we een onderscheid maken tussen betrouwbaarheids- en predictie-intervallen.

1.5.2 Betrouwbaarheidsintervallen

Om betrouwbaarheidsintervallen voor predicties te kunnen opstellen maken we gebruik van distributionele assumpties en daarom is een extra assumptie m.b.t. de fouttermen noodzakelijk:

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

(i.i.d.= onafhankelijk en identisch verdeeld)
Dit betekent dat de fouttermen ε_i allen onafhankelijk zijn van elkaar en normaal verdeeld zijn. Als gevolg daarvan zijn de schatters voor de regressiecoëfficiënten $\mathbf{B}$ normaal verdeeld met verwachtingswaarde en variantie zoals hier boven aangegeven. σ^2 is echter ongekend en wordt geschat door S^2.

$$\begin{aligned} E(\mathbf{B}) &= \boldsymbol{\beta} \\ \widehat{\mathrm{Var}}(\mathbf{B}) &= S^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Als bovenstaande assumptie m.b.t. de fouttermen geldt, dan geldt ook dat de individuele observaties Y_i normaal verdeeld zijn met verwachtingswaarde $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$ en variantie σ^2.

Een $(1-\alpha) \times 100\%$ betrouwbaarheidsinterval voor $E(Y_*|x_{1*}, x_{2*}, \ldots, x_{p*})$ wordt gegeven door

$$\left[\hat{Y}_* - |t_{n-(p+1);\alpha/2}| S\left\{\hat{Y}_*\right\}, \hat{Y}_* + |t_{n-(p+1);\alpha/2}| S\left\{\hat{Y}_*\right\}\right]$$

met p het aantal predictoren in het model en $S\left\{\hat{Y}_*\right\}$ de standaardfout van het geschatte gemiddelde (de uitdrukking kan afgeleid worden maar komt hier niet aan bod).

Dit interval bevat met kans $1-\alpha$ de gemiddelde waarde voor Y_* (op populatieniveau!), gegeven de set van p predictoren. Op basis van het geschatte regressiemodel kunnen we bijgevolg naast een schatter voor $E(Y_*|x_{1*}, x_{2*}, \ldots, x_{p*})$ ook een interval opstellen waarvan met een bepaalde betrouwbaarheid gesteld kan worden dat het de verwachte waarde bevat.

In R kunnen we naast de predicties ook een bijhorend betrouwbaarheidsinterval opvragen.

```
> predict(fit3_expertise,newdata=avprofiel,interval='confidence')
    fit       lwr       upr
1 0.3080308 0.2905969 0.3254647
```

Onder `lwr` lezen we de ondergrens van het interval af, onder `upr` de bovengrens. Standaard krijgen we een 95% betrouwbaarheidsinterval. Het betrouwbaarheidsniveau kan ook aangepast worden, bvb. een 90% betrouwbaarheidsinterval bekomen we als volgt:

```
> predict(fit3_expertise,newdata=avprofiel,interval='confidence',level=0.90)
    fit       lwr       upr
1 0.3080308 0.2934209 0.3226407
```

1.5.3 Predictie-intervallen

Bij het opstellen van predictie-intervallen wordt ook de assumptie van normaal verdeelde fouttermen gemaakt:

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

Een $(1-\alpha) \times 100\%$ predictie-interval voor Y_* wordt gegeven door

$$\left[\hat{Y}_* - |t_{n-(p+1);\alpha/2}| S_{\text{pred}}\left\{\hat{Y}_*\right\}, \hat{Y}_* + |t_{n-(p+1);\alpha/2}| S_{\text{pred}}\left\{\hat{Y}_*\right\}\right]$$

met $S_{\text{pred}}\left\{\hat{Y}_*\right\}$ de standaardfout van de predictie. We moeten hier niet enkel de onzekerheid van de geschatte regressiecoëfficiënten in rekening brengen maar ook de spreiding op de individuele observaties. Bijgevolg is $S_{\text{pred}}\left\{\hat{Y}_*\right\} > S\left\{\hat{Y}_*\right\}$ en is een predictie-interval breder dan een betrouwbaarheidsinterval.

Een $(1-\alpha)\times 100\%$ predictie-interval bevat met kans $1-\alpha$ de (toekomstige) uitkomst Y_* (gegeven de set van p predictoren).

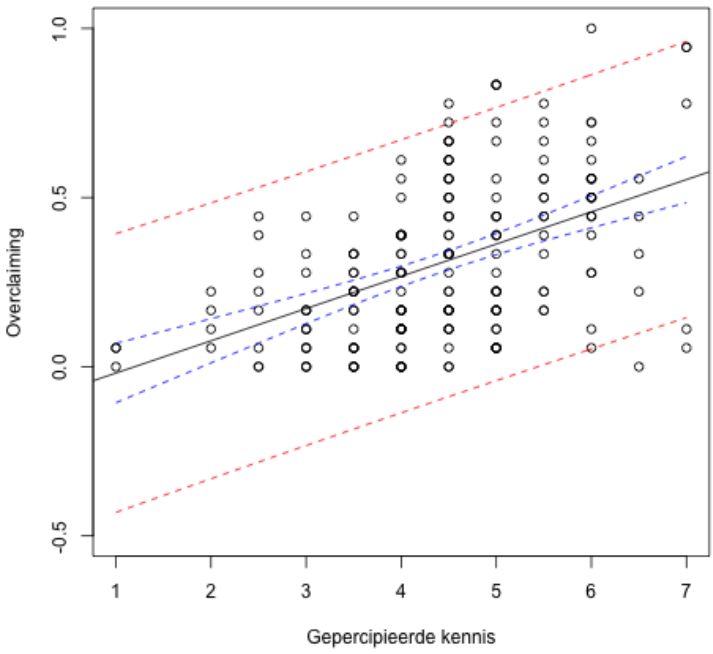
Voor het voorbeeld rond overclaiming (zie sectie 1.2.1) krijgen we volgend 95% predictie-interval voor een 'typisch' profiel:

```
> predict(fit3_expertise,newdata=avprofiel,interval='prediction')
    fit        lwr       upr
1 0.3080308 0.05963625 0.5564254
```

Een 99% predictie-interval bekomen we als volgt:

```
> predict(fit3_expertise,newdata=avprofiel,interval='prediction',level=0.99)
    fit         lwr       upr
1 0.3080308 -0.01957595 0.6356376
```

Om grafisch het verschil tussen een betrouwbaarheidsinterval en predictie-interval te demonstreren, kijken we naar het enkelvoudig regressiemodel waarbij overclaiming geregresseerd wordt op gepercipieerde kennis. De figuur toont voor elke waarde van gepercipieerde kennis de voorspelde mate van overclaiming (i.e. de regressierechte), met bijbehorend betrouwbaarheidsinterval (blauw) en predictie-interval (rood).



Voor elke waarde van gepercipieerde kennis is het betrouwbaarheidsinterval inderdaad smaller dan het predictie-interval. Verder zien we dat het betrouwbaarheidsinterval smaller is in het midden: de standaardfouten van de predicties zijn het kleinst in het midden van de puntenwolk.

1.6 Modelassumpties en invloedrijke observaties

Om de assumpties van het regressiemodel na te gaan kunnen diagnostische plots van de geobserveerde residuen gemaakt worden.

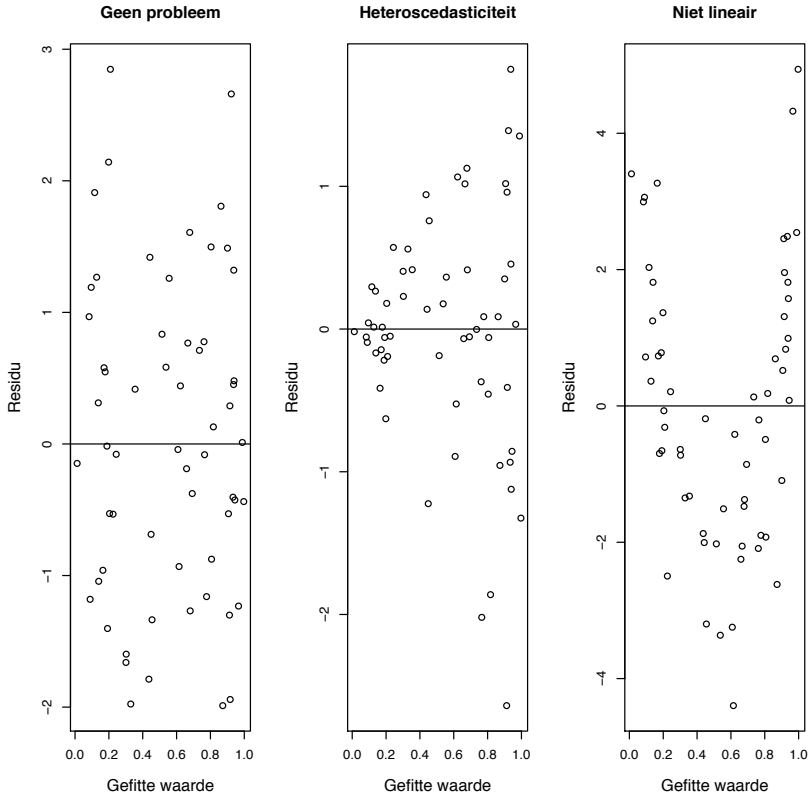
Eén van de belangrijkste plots is een spreidingsdiagram van de geobserveerde residuen e_i in functie van de gefitte waarden $\hat{y}_i$. Op deze plot kunnen we nagaan of

- er sprake is van heteroscedasticiteit (niet-constante variantie van de residuen).
- er sprake is van een niet-lineaire relatie tussen de predictoren en de uitkomst.

  Dit is het geval als er op de figuur duidelijk een trend waar te nemen is, bvb. een kwadratische trend.

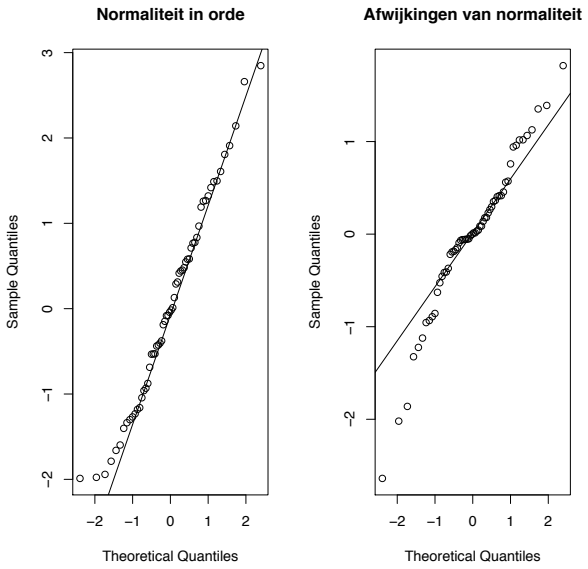
Dit kan men verhelpen door het structurele deel van het model aan te passen, bvb. door ook kwadraten van sommige predictoren in het model op te nemen.

Als alles in orde is, liggen de residuen symmetrisch rond 0 gespreid. De onderstaande figuur illustreert de verschillende gevallen.



Verder moeten we ook nagaan of de residuen normaal verdeeld zijn (indien we gebruik maken van intervalschattingen en/of toetsen):

- Maak een histogram of een boxplot van de gestandaardiseerde residuen, hiermee kan de symmetrie van de verdeling bekeken worden.
- Maak een normale QQ-plot (quantile-quantile plot) van de gestandaardiseerde residuen. Op deze figuur worden de geobserveerde kwantielen t.o.v. de verwachte kwantielen onder de normale verdeling geplot. Systematische afwijkingen van de rechte betekenen afwijkingen van de normaliteit.



In R kunnen diagnostische plots op een eenvoudige manier verkregen worden via de functie `plot` van het object dat het gefitte lineair regressiemodel omvat. Herneem het voorbeeld rond overclaiming (zie sectie 1.2.1).

```
> fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
                    +FINRA_score,data=expertise)
> plot(fit3_expertise)
```

Dit commando geeft ons 4 verschillende plots:

Residuals vs fitted Zie hierboven. Op de plot geeft een rode lijn de geobserveerde trend in de residuen weer. We verwachten een horizontale lijn (op 0).

Normal QQ QQ-plot van de residuen.

Scale-Location Deze plot toont de vierkantswortel van de absolute waarde van de gestandaardiseerde residuen in functie van de gefitte waarden. Een rode lijn geeft de geobserveerde trend weer. In het geval van homoscedasticiteit verwachten we een horizontale lijn (niet noodzakelijk rond 0!).

Residuals vs Leverage De plot toont de gestandaardiseerde residuen t.o.v. de *leverage*. Deze plot laat toe om invloedrijke observaties te detecteren. Dit zijn observaties die een (grote) invloed hebben op de geschatte regressievergelijking.

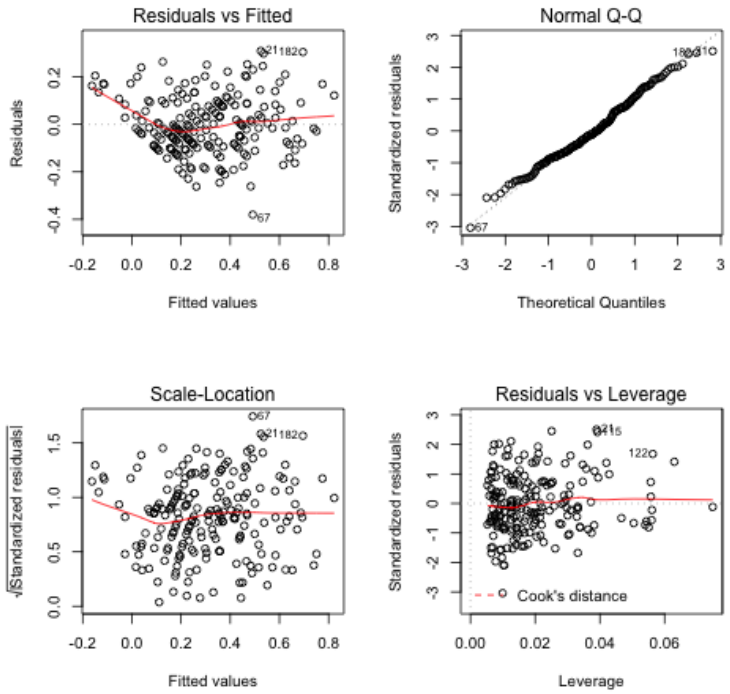
Observaties met een hoge leverage zijn observaties met extreme waarden voor de predictor(en). In het geval van p predictoren is de leverage een afstandsmaat tussen de vector van de p geobserveerde scores voor observatie i en de vector met de gemiddeldes over alle observaties. Een vuistregel stelt dat de leverage groot is als ze groter is dan $2p/n$ met n het aantal observaties. Een punt met extreme waarden voor de predictoren heeft potentieel een grote invloed, zelfs al ligt het niet zo ver van de regressielijn of -vlak; het is immers best mogelijk dat dat punt de regressielijn- of vlak al sterk naar zich toe getrokken heeft.

Het is ook mogelijk dat punten met extreme waarden voor de predictoren toch weinig invloed hebben op de geschatte regressievergelijking. Om de invloed van een welbepaalde observatie na te gaan, kan men de regressie uitvoeren met en zonder die observatie. Cook's distance biedt inzicht in de impact van een welbepaalde observatie i op de predictie van alle andere punten (en niet louter op de predictie van de observatie zelf). Een vuistregel luidt om waarden vanaf 0.8 (voor kleine n) en zeker vanaf 1 als een aanduiding van sterke invloed te zien.

Op deze plot zijn observaties buiten een rode gestreepte lijn observaties met een hoge Cook's distance. Resultaten kunnen drastisch wijzigen wanneer deze observaties uit de analyse weggelaten worden.

Merk op dat R op bovenstaande plots soms het observatienummer naast een punt zet. Dit punt kan dan als een outlier of afwijkende waarde beschouwd worden (maar dit is niet noodzakelijk een invloedrijke observatie). Het loont de moeite om in sommige gevallen deze observaties in meer detail te bekijken.

Onderstaande figuur toont de output voor het regressiemodel `fit3_expertise` (overclaiming):



2 De grootte van een effect

Het is niet ongebruikelijk bij lineaire regressie dat de focus ligt op het nagaan of de effecten van de predictoren op de uitkomst statistisch significant zijn (zie verder bij toetsing). Echter, statistische significantie is niet equivalent met praktische significantie.

Bij het rapporteren van resultaten is het noodzakelijk om naast de resultaten van een toets ook de grootte van effecten mee te geven. We beschouwen een aantal mogelijke r-maten (associatiematen) hiervoor alsook een betrouwbaarheidsinterval voor de regressiecoëfficiënten. In de praktijk kunnen beiden naast elkaar gerapporteerd worden, ze verschaffen andere informatie.

2.1 De determinatiecoëfficiënt R^2

De determinatiecoëfficiënt laat toe om te kwantificeren hoeveel van de variatie in de uitkomst verklaard wordt door het regressiemodel.

Als we denken aan een enkelvoudig regressiemodel waarbij Y op X geregresseerd wordt, weten we reeds dat de lineaire regressielijn de tendens van de lineaire relatie aangeeft. We kunnen echter niet verwachten dat de punten van de observaties op de scatterplot perfect op deze lijn liggen. Analoog zijn bij meervoudige lineaire regressie de predicties $\hat{Y}$ typisch niet perfect gelijk aan de geobserveerde uitkomsten. Dit impliceert dat de predictoren in een regressiemodel de uitkomst Y niet volledig verklaren.

Een maat voor de sterkte van de lineaire regressie:

- is de bekwaamheid van de onafhankelijke variabelen om de variantie van de afhankelijke variabele Y te verklaren.
- wordt bepaald door de grootte van de afwijkingen van de observaties tot de predicties.

  Bij enkelvoudige lineaire regressie wordt de lineaire relatie tussen predictor X en uitkomst Y sterker naarmate de observaties zich dichter bij de regressielijn bevinden.

De totale variatie in Y kan opgesplitst worden in 2 delen: het gedeelte dat verklaard wordt door de regressievergelijking en het gedeelte dat niet kan verklaard worden door de regressievergelijking. De variatie in Y wordt bepaald door de *totale kwadratensom (total sum of squares)* SST:

$$\text{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Concreet kan men aantonen dat:

$$\underbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{\text{SSE}}.$$

SSR is de **regressie kwadratensom (regression sum of squares)** en stelt het deel voor van de variatie in Y dat verklaard wordt door de lineaire regressie van Y op de predictoren[3]. SSE, de **fout kwadratensom** (zie ook hierboven), is het deel van de variatie in Y dat niet verklaard kan worden door de regressie, bij enkelvoudige lineaire regressie is dit de spreiding van de punten rond de regressielijn.

[3] De regressie kwadratensom stelt het deel van de variatie in Y voor dat verklaard wordt door het regressiemodel. SSR wordt daarom ook vaak genoteerd als SS_{Model} of SS_{Mod}.

De determinatiecoëfficiënt R^2 wordt gegeven door de verhouding van de verklaarde variatie op de totale variatie:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST-SSE}}{\text{SST}}$$

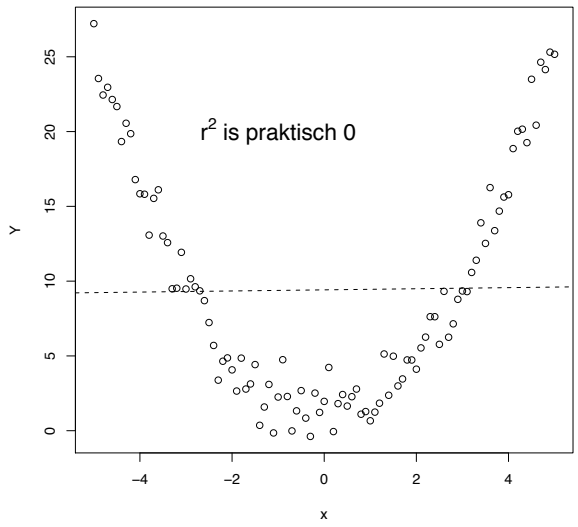
Interpretatie:

- R^2 is de proportie van de totale variatie in de uitkomst Y die verklaard wordt door de predictoren in de lineaire regressie van Y op de predictoren, $0 \leq R^2 \leq 1$.
- R^2 is het kwadraat van de steekproefcorrelatie $\text{Cor}(Y, \hat{Y})$, $-1 \leq R \leq 1$.
- Als $R^2 = 1 (= 100\%)$ dan is SST = SSR m.a.w. SSE = 0, dit wil zeggen dat alle residuen e_i gelijk zijn aan 0. De predicties op basis van het model komen dan perfect overeen met de observaties.
- Als $R^2 = 0 (= 0\%)$ dan is SSR = 0 m.a.w. geen enkel stukje van de variatie in Y wordt door de regressie verklaard. De predictoren zoals opgenomen in het lineair regressiemodel hebben dus geen enkele invloed bij het verklaren van de variatie van Y.

Bij enkelvoudige lineaire regressie met 1 predictor X wordt R^2 soms voorgesteld als r^2 aangezien in dat geval $r = \text{cor}(X, Y)$.

Let op:

- Een hoge R^2 betekent niet noodzakelijk dat nuttige predicties gemaakt kunnen worden.
  R^2 zegt niets over de precisie waarmee predicties gemaakt worden.
- Een hoge R^2 impliceert niet automatisch dat de geschatte regressievergelijking een goede fit is voor de data.
  Bij modellen met een hoge R^2 is het bvb. mogelijk dat modellen waarbij een niet-lineair verband tussen predictoren en uitkomst verondersteld wordt, een betere fit zijn voor de data.
- Een R^2 die dicht bij 0 ligt betekent niet automatisch dat er geen verband is tussen de uitkomst en de set van predictoren.
  R^2 geeft enkel de sterkte van het *lineaire* verband tussen Y en de (lineaire) combinatie van de predictoren weer.



R^2 stijgt naarmate men meer predictoren of onafhankelijke variabelen in het model toevoegt; bovendien is R^2 een vertekende schatter van de ware determinatiecoëfficiënt in de populatie. Om te corrigeren voor het aantal predictoren in het model en een meer onvertekende schatter te bekomen, maakt men in de praktijk vaak gebruik van de aangepaste (Engels: 'adjusted') meervoudige determinatiecoëfficiënt R_a^2:

$$R_a^2 = 1 - \frac{n-1}{n-(p+1)}(1-R^2)$$

Merk op dat R_a^2 steeds kleiner is dan R^2 en niet steeds positief. Merk ook op dat R_a^2 niet dezelfde betekenis heeft als R^2, wees voorzichtig bij het interpreteren van deze statistiek.

In R wordt via de de `summary` van het gefitte model R^2 en R_a^2 van het model weergegeven.

Wanneer we in het voorbeeld rond overclaiming (zie sectie 1.2.1) de uitkomst overclaiming regresseren op gepercipieerde kennis, accuraatheid en FINRA-score, bekomen we:

```
> summary(fit3_expertise)

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge +
accuracy + FINRA_score, data = expertise)
```

```
Residuals:
Min       1Q   Median       3Q      Max
-0.38033 -0.08672 -0.01418  0.08808  0.30886

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.057787   0.039203   1.474   0.1421
self_perceived_knowledge  0.094069   0.008018  11.732   <2e-16 ***
accuracy                 -0.793219   0.045655 -17.374   <2e-16 ***
FINRA_score               0.018370   0.008576   2.142   0.0334 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1256 on 198 degrees of freedom
Multiple R-squared:  0.7116,Adjusted R-squared:  0.7073
F-statistic: 162.9 on 3 and 198 DF,  p-value: < 2.2e-16
```

We zien dat $R^2 = 0.7116$ (`Multiple R-squared`) wat betekent dat 71.16% van de variatie in overclaiming verklaard wordt door het model. $R^2_a = 0.7073$ (`Adjusted R-squared`).

We kunnen R^2 uit bovenstaande output als volgt bekomen:

```
> summary(fit3_expertise)$r.squared
[1] 0.7116225
```

2.2 Semi-partiële en partiële correlatie

R^2 geeft de proportie variatie in de uitkomst verklaard door het volledige model. Partiële maten laten toe om de bijdrage van de afzonderlijke predictoren te kwantificeren. We beschouwen hier 2 r-maten die de samenhang tussen de uitkomst en een individuele predictor geven, conditioneel op de overige predictoren (i.e. gegeven dat deze constant gehouden worden).

Laat $R^2_{Y|x_1,x_2,\ldots,x_p}$ de determinatiecoëfficiënt voorstellen van de regressie van Y op de p predictoren X_1, X_2, $\ldots$, X_p.

Analoog stelt $R^2_{Y|x_1,\ldots,x_{\ell-1},x_{\ell+1},\ldots,x_p}$ de determinatiecoëfficiënt voor van de regressie met dezelfde predictoren maar zonder X_ℓ in het model.

De **semi-partiële correlatie van** X_ℓ met Y is gelijk aan

$$sr_\ell = \sqrt{R^2_{Y|x_1,x_2,\ldots,x_p} - R^2_{Y|x_1,\ldots,x_{\ell-1},x_{\ell+1},\ldots,x_p}}.$$

sr_ℓ^2 is het gedeelte van de variatie van Y dat X_ℓ toelaat te verklaren bovenop het gedeelte dat door de andere $p-1$ predictoren voorspeld kan worden.

De **partiële correlatie** van X_ℓ met Y is gelijk aan

$$pr_\ell = \sqrt{\frac{sr_\ell^2}{1 - R^2_{Y|x_1,\ldots,x_{\ell-1},x_{\ell+1},\ldots,x_p}}}.$$

Dit geeft het verband weer tussen Y en X_ℓ nadat beide uitgezuiverd zijn voor het gedeelte dat met $X_1, \ldots, X_{\ell-1}, X_{\ell+1}, \ldots, X_p$ samenhangt.

Merk op dat een vierkantswortel zowel positief als negatief kan zijn; hier is het **teken** van sr_ℓ en pr_ℓ gelijk aan het teken van de geschatte regressiecoëfficiënt van predictor X_ℓ!

Aangezien β_ℓ de verwachte verandering in Y weergeeft indien X_ℓ met één eenheid stijgt terwijl de andere predictoren constant blijven, wordt hiernaar soms verwezen met de term **partiële regressiecoëfficiënt**.

Een andere (maar in dit geval equivalente) manier om sr_ℓ en pr_ℓ te bekomen is via kwadratensommen. Laat SSR en SSE respectievelijk de verklaarde kwadratensom en fout kwadratensom voorstellen van het model waarin ook predictor X_ℓ opgenomen is en SSR_ℓ de verklaarde kwadratensom van het model zonder X_ℓ. De kwadratensom die bij predictor X_ℓ hoort is dan gelijk aan $\text{SSR} - \text{SSR}_\ell$ en stellen we voor door SS_{X_ℓ}. Er geldt dat

$$\begin{aligned} sr_\ell^2 &= \frac{\text{SS}_{X_\ell}}{SST} \\ pr_\ell^2 &= \frac{\text{SS}_{X_\ell}}{\text{SS}_{X_\ell} + SSE} \end{aligned}$$

In R kan gebruik maakt worden van het package `lsr` om op een eenvoudige manier de semi-partiële en partiële correlatie te bekomen. We gebruiken hiervoor het commando `etaSquared`. We specifiëren hierbij dat we wensen gebruik te maken van Type III kwadratensommen (i.e. de kwadratensommen die we beschouwen in deze cursus).

Voor het voorbeeld rond overclaiming (zie sectie 1.2.1) krijgen we:

```
> etaSquared(fit3_expertise,type=3)
                              eta.sq    eta.sq.part
self_perceived_knowledge 0.200467989  0.41008454
accuracy                 0.439646003  0.60388983
FINRA_score              0.006681956  0.02264613
```

In de kolom `eta.sq` lezen we voor iedere predictor sr^2_ℓ af. Zo kunnen we afleiden dat de semi-partiële correlatie tussen FINRA-score en de uitkomst gelijk is aan $\sqrt{0.00668} = 0.0817$. In de kolom `eta.sq.part` lezen we pr^2_ℓ. De partiële correlatie tussen FINRA-score en de uitkomst is bijgevolg gelijk aan $\sqrt{0.0226} = 0.15$. Merk op dat de correlaties positief zijn aangezien het teken van de geschatte regressiecoëfficiënt voor de FINRA-score positief is.

De output kan ook verkregen worden samen met de kwadratensommen:

```
> etaSquared(fit3_expertise,type=3,anova=TRUE)
                             eta.sq eta.sq.part       SS    df         MS
self_perceived_knowledge 0.200467989  0.41008454 2.17303527   1 2.17303527
accuracy                 0.439646003  0.60388983 4.76567991   1 4.76567991
FINRA_score              0.006681956  0.02264613 0.07243114   1 0.07243114
Residuals                0.288377524          NA 3.12595808 198 0.01578767
                                F          p
self_perceived_knowledge 137.641316 0.00000000
accuracy                 301.860933 0.00000000
FINRA_score                4.587831 0.03342126
Residuals                       NA         NA
```

De waarden voor de verschillende kwadratensommen lezen we af onder `SS` (we concentreren ons hier enkel op dit stuk van de output). De fout kwadratensom kunnen we aflezen bij `Residuals`, SSE=3.12596. De totale kwadratensom SST kunnen we in R berekenen:

```
> sst<-sum((expertise$overclaiming_proportion-mean(expertise$overclaiming_proportion))^2)
> sst
[1] 10.83981
```

Voor FINRA-score lezen we af dat

- $sr^2_\ell = \frac{0.07243}{10.8398} = 0.006682$ en
- $pr^2_\ell = \frac{0.07243}{0.07243+3.1260} = 0.02265$.

Merk op dat bij `eta.sq` bij de `Residuals` SSE/SST weergegeven wordt, dit is dus $1 - R^2$ (de proportie van de variatie in de uitkomst die niet verklaard wordt door het model).

2.3 Betrouwbaarheidsintervallen voor $\boldsymbol{\beta}$

Om betrouwbaarheidsintervallen voor de regressiecoëfficiënten te kunnen opstellen wordt ook de assumptie van normaal verdeelde fouttermen gemaakt:

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

Een betrouwbaarheidsinterval voor een parameter heeft typisch de volgende vorm:

$$\text{schatter} \pm \text{ (kritische waarde) } \times \text{(Standaardfout van schatter)}$$

Onder de bovenstaande assumptie m.b.t. de fouttermen, kan aangetoond worden dat een $(1-\alpha)\times 100\%$ betrouwbaarheidsinterval voor β_ℓ $(\ell = 1, \ldots, p)$ er als volgt uit ziet:

$$\left[B_\ell - |t_{n-(p+1);\alpha/2}| S\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{\ell,\ell}}, B_\ell + |t_{n-(p+1);\alpha/2}| S\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{\ell,\ell}}\right]$$

waarbij p het aantal predictoren in het model is en $(\mathbf{X}'\mathbf{X})_{\ell,\ell}$ het element op rij ℓ en kolom ℓ van de matrix $(\mathbf{X}'\mathbf{X})^{-1}$ voorstelt. $S\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{\ell,\ell}}$ is bijgevolg de standaardfout van B_ℓ.

Hoewel de steekproevenverdeling van B_ℓ een normale verdeling is, moeten we gebruik maken van de kristische waarde $t_{n-(p+1);\alpha/2}$ uit de t-verdeling met $n-(p+1)$ vrijheidsgraden aangezien we moeten corrigeren voor het feit dat σ^2 geschat wordt door S^2.

Dit betrouwbaarheidsinterval omvat β_ℓ met kans $1-\alpha$.

Deze intervallen geven ons naast de grootte van een effect informatie over de nauwkeurigheid van de schatter.

Wanneer we in het voorbeeld rond overclaiming (zie sectie 1.2.1), de uitkomst overclaiming regresseren op gepercipieerde kennis, accuraatheid en FINRA-score, kunnen we als volgt een 95% betrouwbaarheidsinterval voor het effect van elke predictor bekomen:

```
> confint(fit3_expertise)
                            2.5 %      97.5 %
(Intercept)             -0.019521646  0.13509525
self_perceived_knowledge  0.078257283  0.10988105
accuracy                -0.883251647 -0.70318621
FINRA_score              0.001457144  0.03528216
```

De geschatte regressiecoëfficiënt die het geschatte effect van gepercipieerde kennis, na correctie voor accuraatheid en FINRA score, op de gemiddelde uitkomst weergeeft, is gelijk aan 0.09407 (zie output hieronder). Een 95% betrouwbaarheidsinterval voor het effect is gelijk aan $[0.0783; 0.1099]$.

```
> fit3_expertise

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge +
accuracy + FINRA_score, data = expertise)

Coefficients:
(Intercept)  self_perceived_knowledge                  accuracy
0.05779                     0.09407                   -0.79322
FINRA_score
0.01837
```

Standaard geeft R via `confint` een 95% betrouwbaarheidsinterval, maar we kunnen eenvoudig ook andere niveaus opvragen, bvb. een 90% betrouwbaarheidsinterval:

```
> confint(fit3_expertise, level=0.90)
                                5 %        95 %
(Intercept)              -0.006999055  0.12257266
self_perceived_knowledge  0.080818526  0.10731981
accuracy                 -0.868667951 -0.71776990
FINRA_score               0.004196669  0.03254263
```

3 Regressie met nominale predictoren

Hoewel de term 'regressie' in de klassieke terminologie inhoudt dat alle predictoren van minstens intervalniveau zijn, is lineaire regressie technisch perfect mogelijk met nominale predictoren. Een variabele van nominaal niveau wordt een **factor** genoemd (dit is ook de terminologie gehanteerd door R).

Wanneer alle predictoren van nominaal niveau zijn, spreekt men van **variantie-analyse**. In sectie 8.2 gaan we daar dieper op in.

3.1 Lineaire regressie met hulpveranderlijken

In het algemeen geldt dat we een nominale predictor met I niveaus moeten hercoderen tot $I - 1$ nieuwe hulpveranderlijken die we vervolgens in het regressiemodel kunnen stoppen.

We kunnen hierbij een onderscheid maken tussen **dummy-codering** en **effect-codering**. We bekijken dit aan de hand van een voorbeeld.

Veronderstel dat we het effect van het type onderwijs op de uiteindelijke studieresultaten wensen te modelleren waarbij er in totaal 4 types van onderwijs beschouwd worden. De variabele die het type onderwijs weergeeft is nominaal.

- Bij **dummy-codering** kiest men één van de I niveaus als referentieniveau en worden de andere niveaus via een 0-1 variabele gecodeerd.

  In het geval van het voorbeeld betekent dit dat we 3 hulpveranderlijken X_1, X_2 en X_3 moeten aanmaken. Wanneer we type 4 als referentieniveau beschouwen, dan bekomen we de volgende codering:

Type onderwijs	X_1	X_2	X_3
Type 1	1	0	0
Type 2	0	1	0
Type 3	0	0	1
Type 4	0	0	0

  Dit betekent concreet dat voor een individu i

  - die type 1 van onderwijs volgt, geldt: $x_{i1} = 1$, $x_{i2} = 0$, $x_{i3} = 0$.
  - die type 2 van onderwijs volgt, geldt: $x_{i1} = 0$, $x_{i2} = 1$, $x_{i3} = 0$.
  - die type 3 van onderwijs volgt, geldt: $x_{i1} = 0$, $x_{i2} = 0$, $x_{i3} = 1$.
  - die type 4 van onderwijs volgt, geldt: $x_{i1} = 0$, $x_{i2} = 0$, $x_{i3} = 0$.

- **Effect-codering** is analoog aan dummy-codering behalve dat de referentiegroep steeds met -1 gecodeerd wordt i.p.v. met 0. Voor het voorbeeld bekomen we:

Type onderwijs	X_1	X_2	X_3
Type 1	1	0	0
Type 2	0	1	0
Type 3	0	0	1
Type 4	-1	-1	-1

  Dit betekent dat de codering hetzelfde is als de dummy-codering voor individuen die type 1, 2 of 3 van het onderwijs volgen maar voor een individu i die type 4 van onderwijs volgt, geldt: $x_{i1} = -1$, $x_{i2} = -1$, $x_{i3} = -1$.

Het effect van het type onderwijs op de verwachtingswaarde van de studieresultaten Y kunnen we als volgt modelleren:

$$E(Y_i|x_{i1}, x_{i2}, x_{i3}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Naargelang het coderingsschema dat gehanteerd wordt, hebben de regressieparameters een andere betekenis.

- Dummy-codering
  - $E(Y_i|\text{Type 4}) = E(Y_i|x_{i1} = 0, x_{i2} = 0, x_{i3} = 0) = \beta_0$. β_0 stelt dus het verwachte studieresultaat voor bij type 4 van onderwijs.
  - $E(Y_i|\text{Type 1}) = E(Y_i|x_{i1} = 1, x_{i2} = 0, x_{i3} = 0) = \beta_0 + \beta_1$. β_1 stelt dus het verschil voor van het verwachte studieresultaat bij type 1 en het verwachte studieresultaat bij type 4 van onderwijs.

    Analoog stellen β_2 en β_3 het verschil in verwachte studieresultaat tussen type 2 en type 4 en tussen type 3 en type 4.
- Effect-codering
  - In dit geval kan aangetoond worden dat β_0 het marginale gemiddelde van het studieresultaat voorstelt, i.e. het gemiddelde van de gemiddelde studieresultaten over de verschillende onderwijstypes heen:
    $(E(Y_i|\text{Type 1}) + E(Y_i|\text{Type 2}) + E(Y_i|\text{Type 3}) + E(Y_i|\text{Type 4}))/4$.
  - β_ℓ ($\ell = 1, 2, 3$) drukt dan het verschil uit tussen het verwachte studieresultaat in onderwijstype ℓ en het marginale gemiddelde.
  - Het verwachte studieresultaat in type 4 van het onderwijs is

    $$E(Y_i|\text{Type 4}) = E(Y_i|x_{i1} = -1, x_{i2} = -1, x_{i3} = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3.$$

    Dit betekent dat het verschil tussen het verwachte resultaat in onderwijstype 4 en het marginale gemiddelde gelijk is aan $\beta_4 = -\beta_1 - \beta_2 - \beta_3$. Bijgevolg geldt dat $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 0$.

3.2 Voorbeeld: pijneducatie

Wanneer we via lineaire regressie het effect van `conditie` (nominaal, 3 niveaus) op de uitkomst `Buig` (verschilscore in voorover buigen) modelleren (zie sectie 1.2.2; we laten de overige variabelen hier buiten beschouwing), vergelijken we de gemiddelde uitkomst over de 3 condities.

Na het inlezen van de data, zien we dat de variabele `conditie` in R als een factor met 3 niveaus gedefinieerd is.

```
> class(pijneducatie$Conditie)
[1] "factor"
> levels(pijneducatie$Conditie)
[1] "Algemene pijneducatie" "Baseline"              "Rugpijneducatie"
```

Via `contrasts()` kunnen we opvragen welke restrictieschema gehanteerd wordt.

```
> contrasts(pijneducatie$Conditie)
                          Baseline Rugpijneducatie
Algemene pijneducatie        0               0
Baseline                     1               0
Rugpijneducatie              0               1
```

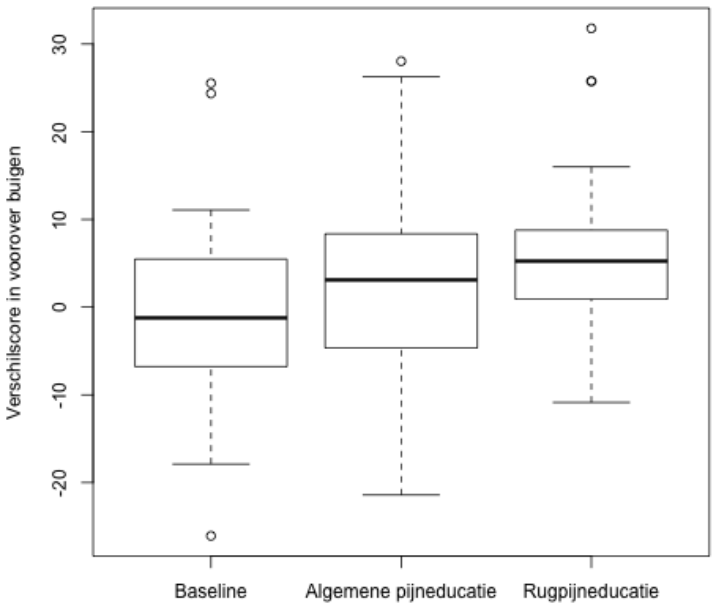
We zien dat dit standaard de dummy-codering is. R zal als referentieniveau altijd het eerste niveau kiezen. Echter, de niveaus werden in dit geval in R alfabetisch gerangschikt wat betekent dat we `conditie=Baseline` en `conditie=Rugpijneducatie` vergelijken met `conditie=Algemene pijneducatie`. Gezien we hier 2 'behandelingen' hebben en 1 baseline, is het logisch om de baseline als referentieniveau te kiezen.

```
pijneducatie$Conditie<-factor(pijneducatie$Conditie,levels=c("Baseline","Algemene
                                    pijneducatie","Rugpijneducatie"))
```

Op die manier geven we aan in R dat het eerste niveau (het referentieniveau) `conditie=Baseline` is.

```
> contrasts(pijneducatie$Conditie)
                          Algemene pijneducatie Rugpijneducatie
Baseline                                    0               0
Algemene pijneducatie                       1               0
Rugpijneducatie                             0               1
```

De verdeling van de uitkomst per groep kan grafisch als volgt voorgesteld worden:



De steekproefgemiddeldes per groep zijn:

```
Baseline          Algemene pijneducatie       Rugpijneducatie
-0.765494               2.608108              5.063635
```

We stellen vast dat de geobserveerde gemiddelde uitkomst het laagst is in de baseline conditie en het hoogst in de conditie met rugpijneducatie.

Het marginale (ongewogen) steekproefgemiddelde is gelijk aan $(-0.765494 + 2.608108 + 5.063635)/3 = 2.30208$.

3.2.1 Dummy-codering voor conditie

We weten reeds dat `conditie` in R als een factor gedefinieerd is en dat standaard de dummy-codering gebruikt zal worden. We hoeven dus verder niets te specifiëren wanneer we `Buig` op `Conditie` regresseren.

```
> fit_pijneducatie_dummy<-lm(Buig~Conditie,data=pijneducatie)
> summary(fit_pijneducatie_dummy)

Call:
lm(formula = Buig ~ Conditie, data = pijneducatie)

Residuals:
Min       1Q   Median       3Q      Max
-25.3164  -5.5974  -0.0914   4.7621  26.7199

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -0.7655     1.5885  -0.482   0.6308
ConditieAlgemene pijneducatie  3.3736     2.2464   1.502   0.1358
ConditieRugpijneducatie        5.8291     2.2327   2.611   0.0102 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.05 on 118 degrees of freedom
Multiple R-squared:  0.05497,Adjusted R-squared:  0.03895
F-statistic: 3.432 on 2 and 118 DF,  p-value: 0.03559
```

We concentreren ons hier op de geschatte regressiecoëfficiënten. We lezen af dat de geschatte gemiddelde uitkomst in de baseline conditie gelijk is aan -0.7655. In de conditie met algemene pijneducatie is dit $-0.7655 + 3.3736 \times 1 + 5.8291 \times 0 = 2.60814$; in de conditie met rugpijneducatie is dit $-0.7655 + 3.3736 \times 0 + 5.8291 \times 1 = 5.0636$. Dit komt overeen met de geobserveerde steekproefgemiddeldes.

Omgekeerd kunnen de geschatte regressiecoëfficiënten afgeleid worden uit de geobserveerde steekproefgemiddeldes.

Het is belangrijk om zeker te zijn dat de nominale variabelen als factoren gedefinieerd zijn en dat men goed weet welk coderingsschema gebruikt wordt. In R staat `contr.treatment` voor dummy-codering. Je kan dit ook zelf instellen voor een factor (dit kan nodig zijn als het coderingsschema niet goed ingesteld staat).

```
> contrasts(pijneducatie$Conditie)<-contr.treatment
> contrasts(pijneducatie$Conditie)
                     2 3
Baseline             0 0
```

```
Algemene pijneducatie  1 0
Rugpijneducatie        0 1
```

Merk op dat bij de naamgeving van de hulpveranderlijken de oorspronkelijke labels in dit geval niet overgenomen worden. Uit het coderingsschema kunnen we afleiden dat `Conditie=2` overeen komt met de algemene pijneducatie en `Conditie=3` met de rugpijneducatie.

```
> fit2_pijneducatie_dummy<-lm(Buig~Conditie,data=pijneducatie)
> summary(fit2_pijneducatie_dummy)

Call:
lm(formula = Buig ~ Conditie, data = pijneducatie)

Residuals:
Min       1Q   Median       3Q      Max
-25.3164  -5.5974  -0.0914   4.7621  26.7199

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7655     1.5885  -0.482   0.6308
Conditie2     3.3736     2.2464   1.502   0.1358
Conditie3     5.8291     2.2327   2.611   0.0102 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.05 on 118 degrees of freedom
Multiple R-squared:  0.05497,Adjusted R-squared:  0.03895
F-statistic: 3.432 on 2 and 118 DF,  p-value: 0.03559
```

Het is, zeker wanneer er meerdere nominale predictoren zijn, handig en veilig om coderingsschema's rechtstreeks via de functie `lm` mee te geven.

```
lm(Buig~Conditie,data=pijneducatie,contrasts=list(Conditie=contr.treatment))
```

3.2.2 Effect-codering voor conditie

Wanneer we effect-codering willen hanteren, doen we dit via `contr.sum` in R.

```
> contrasts(pijneducatie$Conditie)<-contr.sum
> contrasts(pijneducatie$Conditie)
                       [,1] [,2]
Baseline                 1    0
Algemene pijneducatie    0    1
Rugpijneducatie         -1   -1
```

We zien dat beide hulpveranderlijken op -1 gezet worden in de rugpijneducatie.

We kunnen het coderingsschema ook meegeven in de functie `lm`.

```
> fit_pijneducatie_effect<-lm(Buig~Conditie,data=pijneducatie,
                              contrasts=list(Conditie=contr.sum))
> summary(fit_pijneducatie_effect)

Call:
lm(formula = Buig ~ Conditie, data = pijneducatie,
                 contrasts = list(Conditie = contr.sum))

Residuals:
Min       1Q   Median       3Q      Max
-25.3164  -5.5974  -0.0914   4.7621  26.7199

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3021     0.9134   2.520   0.0131 *
Conditie1    -3.0676     1.2943  -2.370   0.0194 *
Conditie2     0.3060     1.2943   0.236   0.8135
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.05 on 118 degrees of freedom
Multiple R-squared:  0.05497,Adjusted R-squared:  0.03895
F-statistic: 3.432 on 2 and 118 DF,  p-value: 0.03559
```

We zien dat de schatting voor het intercept inderdaad gelijk is aan het marginale (ongewogen) steekproefgemiddelde van de gemiddelde uitkomst binnen de 3 condities. Verder leiden we af dat het geschatte gemiddelde binnen de conditie met algemene pijneducatie (`conditie=2`) gelijk is aan $2.3021 - 3.0676 \times 0 + 0.3060 \times 1 = 2.6081$ en dat het geschatte gemiddelde in de

baseline conditie (`conditie=1`) gelijk is aan $2.3021 - 3.0676 \times 1 + 0.3060 \times 0 = -0.7655$. Dit komt overeen met de geobserveerde steekproefgemiddeldes.

Voor de conditie met rugpijneducatie kunnen we afleiden dat het geschatte gemiddelde gelijk is aan $2.3021 - 3.0676 \times (-1) + 0.3060 \times (-1) = 5.0637$.

4 Toetsing

In dit stuk gaan we dieper in op hypothesetoetsen voor de parameters in een regressiemodel. Hierbij wordt ook de assumptie van normaal verdeelde fouttermen gemaakt:

$$\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

4.1 Modelvergelijkingen

Wanneer we meerdere (p) predictoren in het regressiemodel hebben, kunnen we ons afvragen of het nodig is die allemaal in het model op te nemen.

Om dit te toetsen kunnen we het model met p predictoren vergelijken met een model met k predictoren die een subset zijn van de p predictoren.

Dit is een speciaal geval van een algemene toets die nagaat of twee lineaire modellen, met het ene model genest in het andere, significant van elkaar verschillen wat betreft de mogelijkheid om de uitkomst Y te voorspellen.

Een lineair model B is genest in model A wanneer bij model B bijkomende lineaire restricties m.b.t. de te schatten parameters opgelegd worden.

Als we de 2 modellen A (zonder restricties) en B (met restricties) met elkaar wensen te vergelijken, kunnen we gebruik maken van deze toetsingsgrootheid:

$$F = \frac{(\text{SSE}_B - \text{SSE}_A)/(\text{df}_B - \text{df}_A)}{\text{SSE}_A/\text{df}_A} \tag{2}$$

SSE_A stelt de fout kwadratensom van model A voor en SSE_B de fout kwadratensom van model B. df_A en df_B stellen de overeenkomstige vrijheidsgraden in respectievelijk model A en model B voor.

Met p predictoren in het model dienen op basis van n observaties $p + 1$ coëfficiënten geschat te worden (vergeet het intercept niet). Dit betekent dat in dit geval $\text{df}_A = n - (p + 1)$.

Wanneer model B slechts een subset van k predictoren bevat, is $\text{df}_B = n - (k + 1)$.

Er kan aangetoond worden dat onder H_0 (beide modellen zijn niet verschillend van elkaar), de toetsingsgrootheid (2) F-verdeeld is met $\text{df}_B - \text{df}_A$ vrijheidsgraden voor de teller en df_A vrijheidsgraden voor de noemer ($F \sim F(\text{df}_B - \text{df}_A, \text{df}_A)$). Dit is de nuldistributie van F in (2), dit betekent dat we de bijhorende p-waarde van de toets kunnen berekenen en beslissen om H_0 al dan niet te verwerpen.

Dit is een algemene beschrijving voor toetsing aan de hand van modelvergelijkingen. In de volgende secties worden enkele concrete gevallen in meer detail besproken.

4.2 Toets voor alle predictoren

We stellen ons hier de vraag: is er tenminste 1 predictor nuttig in het voorspellen van de uitkomst?

We vergelijken het model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n \tag{3}$$

met het model zonder predictoren:

$$Y_i = \beta_0 + \varepsilon_i.$$

Dit model noemen we het **nulmodel**. Het aantal vrijheidsgraden dat geassocieerd is met het nulmodel is $n - 1$ (er dient enkel een intercept geschat te worden). Voor het nulmodel geldt dat $\hat{Y}_i = B_0 = \bar{Y}$ $(i = 1, \ldots, n)$.

Of nog: we toetsen de volgende nulhypothese

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

tegenover de alternatieve hypothese H_1 die stelt dat tenminste 1 regressiecoëfficiënt verschillend is van 0. Het verwerpen van H_0 betekent dus dat we verder moeten nagaan voor welke predictoren de regressiecoëfficiënt significant verschillend is van 0.

Er geldt dat de fout kwadratensom voor het nulmodel gelijk is aan de totale kwadratensom SST van het volledig model in (3). Het aantal vrijheidsgraden dat overeenkomt met SST is $n - 1$.

Bijgevolg ziet de toetsingsgrootheid in (2) er in dit geval als volgt uit:

$$F = \frac{(\text{SST} - \text{SSE})/(\text{df}_0 - \text{df}_A)}{\text{SSE}/\text{df}_A} = \frac{\text{SSR}/(\text{df}_0 - \text{df}_A)}{\text{SSE}/\text{df}_A}$$

met df_0 het aantal vrijheidsgraden van het nulmodel $(= n - 1)$ en df_A het aantal vrijheidsgraden van het volledige model in (3) $(= n - (p + 1))$. Bijgevolg is $\text{df}_0 - \text{df}_A = p$. Dit is

het aantal vrijheidsgraden dat hoort bij de verklaarde kwadratensom SSR van een model en is gelijk aan het aantal parameters van het volledige model, intercept niet meegerekend.

Beschouw het voorbeeld rond overclaiming (zie sectie 1.2.1). We voorspellen overclaiming op basis van gepercipieerde kennis, accuraatheid en FINRA-score.

```
> fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
             +FINRA_score,data=expertise)
> summary(fit3_expertise)

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge + accuracy
         + FINRA_score, data = expertise)

Residuals:
    Min       1Q   Median       3Q      Max
-0.38033 -0.08672 -0.01418  0.08808  0.30886

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.057787   0.039203   1.474   0.1421
self_perceived_knowledge  0.094069   0.008018  11.732   <2e-16 ***
accuracy                 -0.793219   0.045655 -17.374   <2e-16 ***
FINRA_score               0.018370   0.008576   2.142   0.0334 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1256 on 198 degrees of freedom
Multiple R-squared:  0.7116,Adjusted R-squared:  0.7073
F-statistic: 162.9 on 3 and 198 DF,  p-value: < 2.2e-16
```

In deze output kunnen we op de onderste lijn het resultaat van de F-toets voor alle predictoren aflezen. Hier is de geobserveerde toetsingsgrootheid f^* gelijk aan 162.9, bij de bijhorende p-waarde duidt `< 2.2e-16` aan dat deze p-waarde heel klein is. We hebben sterke evidentie tegen de nulhypothese wat impliceert dat we evidentie hebben voor het feit dat minstens 1 predictor een invloed heeft op de uitkomst. In dit voorbeeld is het aantal observaties $n = 202$. De toetsingsgrootheid volgt onder de nulhypothese een F-verdeling met 3 vrijheidsgraden (aantal parameters dat getoetst wordt) voor de teller en 198 (i.e. 202-(3+1)) voor de noemer.

In dit stukje van de output kunnen we de kwadratensommen zelf niet aflezen. Deze informatie

krijgen we wel als we zelf de modelvergelijkingstoets uitvoeren via het commando `anova`.

Het nulmodel kunnen we als volgt definiëren:

```
fit0_expertise<-lm(overclaiming_proportion~1,data=expertise)
```

Het resultaat van de modelvergelijking die het effect voor alle predictoren toetst is als volgt:

```
> anova(fit0_expertise,fit3_expertise)
Analysis of Variance Table

Model 1: overclaiming_proportion ~ 1
Model 2: overclaiming_proportion ~ self_perceived_knowledge + accuracy +
FINRA_score
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    201 10.840
2    198  3.126  3    7.7139 162.87 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Bovenaan in de output staat weergegeven welke 2 modellen met elkaar vergeleken worden.

In de kolom `RSS` lezen we de fout kwadratensommen af. Voor het nulmodel is deze som gelijk aan 10.840. Dit is ook de SST van het volledige model. Het overeenkomstige aantal vrijheidsgraden kan afgelezen worden in de kolom `Res.Df` en is gelijk aan $n - 1 = 202 - 1 = 201$.

Voor het volledige model is de fout kwadratensom SSE gelijk aan 3.126, het overeenkomstig aantal vrijheidsgraden is gelijk aan $n - (3 + 1) = 202 - 4 = 198$. We lezen op de onderste lijn onder `Sum of Sq` verder af dat SST-SSE=7.7139=SSR en dat het overeenkomstig aantal vrijheidsgraden gelijk is aan 3, i.e. het verschil in aantal vrijheidsgraden tussen het volledige model en het nulmodel of nog, het aantal parameters voor de predictoren in het volledige model, intercept niet inbegrepen.

We zien dat de geobserveerde toetsingsgrootheid inderdaad overeenkomt met wat we voordien bekwamen, namelijk $f^* = (7.7139/3)/(3.126/198) = 162.87$.

4.3 Toets voor een subset van predictoren

De toets voor alle predictoren uit de vorige sectie is een speciaal geval van de modelvergelijkingstoets waarbij een set van predictoren getoetst wordt. Wanneer we in het

voorbeeld rond overclaiming (zie sectie 1.2.1) wensen te toetsen of het model waar naast de gepercipieerde kennis ook accuraatheid en FINRA-score als predictoren opgenomen zijn de variatie in overclaiming beter verklaart, vergelijken we het model met de 3 predictoren met het model met enkel gepercipieerde kennis als predictor.

Als β_2 en β_3 de regressiecoëfficiënten voor respectievelijk accuraatheid en FINRA-score voorstellen, dan toetsen we $H_0 : \beta_2 = \beta_3 = 0$ versus de alternatieve hypothese die stelt dat minstens 1 van beide parameters niet 0 is. De specifiëring in R van de 2 modellen gebeurt als volgt:

```
fit1_expertise<-lm(overclaiming_proportion~self_perceived_knowledge,data=expertise)
fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
                   +FINRA_score,data=expertise)
```

Wanneer we beide modellen met elkaar vergelijken, krijgen we:

```
> anova(fit1_expertise,fit3_expertise)
Analysis of Variance Table

Model 1: overclaiming_proportion ~ self_perceived_knowledge
Model 2: overclaiming_proportion ~ self_perceived_knowledge + accuracy +
FINRA_score
   Res.Df  RSS Df   Sum of Sq    F    Pr(>F)
1    200 8.3303
2    198 3.1260  2    5.2044 164.82 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Model 1 is hier genest in model 2, i.e. in de notatie in (2) is model 1 dus gelijk aan model B terwijl model 2 gelijk is aan model A.

Het aantal vrijheidsgraden (`Res.Df`) voor model 1 is 200 (df_B), i.e. 202-(1+1) terwijl dit 198 (df_A) voor model 2 is.

In kolom `RSS` zien we dat $\text{SSE}_B = 8.3303$ en $\text{SSE}_A = 3.1260$.

Op de onderste lijn zien we verder dat er $200 - 198 = 2 = \text{df}_B - \text{df}_A$ parameters getoetst worden.

De geobserveerde toetsingsgrootheid is $f^* = (5.2044/2)/(3.1260/198) = 164.82$. Onder H_0 is deze toetsingsgrootheid F-verdeeld met 2 en 198 vrijheidsgraden. De bijhorende p-waarde

(`Pr(>F)`) is heel klein waardoor we evidentie hebben tegen de nulhypothese en aannemen dat model 2 de variatie in overclaiming beter verklaart dan model 1.

4.4 Toets voor 1 predictor

We stellen ons hier de vraag: kan 1 welbepaalde predictor uit het model weggelaten worden?

Hiervoor kunnen we opnieuw een modelvergelijkingstoets uitvoeren: het volledige model wordt vergeleken met het model zonder de predictor.

Als deze predictor van intervalniveau is, komt dit overeen met het toetsen van 1 enkele parameter.

Als de predictor van nominaal niveau is, zal het aantal parameters dat getoetst wordt overeenkomen met het aantal hulpveranderlijken voor deze predictor. M.a.w. als we willen nagaan of een nominale predictor een invloed heeft op de uitkomst, vergelijken we de modellen met en zonder de hulpveranderlijken die coderen voor de nominale predictor. Het resultaat van deze toets zal onafhankelijk zijn van het coderingsschema (dummy- of effect-codering) dat gebruikt wordt voor de nominale predictor.

4.4.1 Predictor van intervalniveau

Beschouw het volgende model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_\ell x_{i\ell} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n$$

Wanneer we wensen te toetsen of predictor ℓ uit het model weggelaten kan worden, toetsen we $H_0 : \beta_\ell = 0$ tegenover $H_1 : \beta_\ell \neq 0$.

In het geval dat er slechts 1 parameter getoetst wordt, is de F-toets uit de modelvergelijking equivalent aan een t-toets. Hiervoor kunnen we gebruik maken van de volgende toetsingsgrootheid:

$$T = \frac{B_\ell}{S_{B_\ell}}$$

waarbij S_{B_ℓ} de standaardfout van B_ℓ voorstelt. Er kan aangetoond worden dat onder H_0, $T \sim t_{n-(p+1)}$ of nog: dat T onder H_0 een t-verdeling volgt met $n - (p + 1)$ vrijheidsgraden.

Onder H_0 is de absolute waarde van de toetsingsgrootheid klein (dicht bij 0). Als H_1 geldt, zal de absolute waarde groot zijn.

Op basis van de geobserveerde toetsingsgrootheid $t^* = b_\ell / s_{B_\ell}$ kunnen we de p-waarde berekenen:

$$2 \times P\left(T \geq \left|\frac{b_\ell}{s_{B_\ell}}\right|\right) \text{ met } T \sim t_{n-(p+1)}.$$

Via de output van `lm` in R krijgen we standaard het resultaat van de t-toets voor alle parameters in het model (ook voor het intercept, maar hieraan wordt in de praktijk doorgaans geen aandacht aan besteed).

Herneem het voorbeeld rond overclaiming (zie sectie 1.2.1).

```
> fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
+FINRA_score,data=expertise)
> summary(fit3_expertise)

Call:
lm(formula = overclaiming_proportion ~ self_perceived_knowledge + accuracy
+ FINRA_score, data = expertise)

Residuals:
Min       1Q   Median       3Q      Max
-0.38033 -0.08672 -0.01418  0.08808  0.30886

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.057787   0.039203   1.474   0.1421
self_perceived_knowledge  0.094069   0.008018  11.732   <2e-16 ***
accuracy                 -0.793219   0.045655 -17.374   <2e-16 ***
FINRA_score               0.018370   0.008576   2.142   0.0334 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1256 on 198 degrees of freedom
Multiple R-squared:  0.7116,Adjusted R-squared:  0.7073
F-statistic: 162.9 on 3 and 198 DF,  p-value: < 2.2e-16
```

Voor iedere predictor lezen we in de kolom `t value` de geobserveerde toetsingsgrootheid af en in de kolom `Pr(>|t|)` de overeenkomstige p-waarde.

Zo zien we bvb. voor FINRA-score dat de geobserveerde toetsingsgrootheid $t^* = 0.018370/0.008576 = 2.142$. De bijhorende p-waarde is 0.0334, dit is kleiner dan 5% wat

betekent dat we de nulhypothese dat FINRA-score geen effect heeft op overclaiming kunnen verwerpen op het 5% significantieniveau.

Via `anova` waarbij we de modelvergelijking zelf definiëren komen we tot hetzelfde resultaat:

```
fit2_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy,
                   data=expertise)
fit3_expertise<-lm(overclaiming_proportion~self_perceived_knowledge+accuracy
                  +FINRA_score,data=expertise)
> anova(fit2_expertise,fit3_expertise)
Analysis of Variance Table

Model 1: overclaiming_proportion ~ self_perceived_knowledge + accuracy
Model 2: overclaiming_proportion ~ self_perceived_knowledge + accuracy +
FINRA_score
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    199 3.1984
2    198 3.1260  1  0.072431 4.5878 0.03342 *
```

Er geldt dat $f^* = t^{*2} = (2.142)^2 = 4.59$.

4.4.2 Predictor van nominaal niveau

Om het effect van een nominale predictor met I niveaus te toetsen, voeren we een modelvergelijkingstoets uit met als nulhypothese dat de regressiecoëfficiënten die horen bij de $(I-1)$ hulpveranderlijken allen 0 zijn.

Dit komt neer op toetsen of de gemiddelde uitkomst gelijk is over de verschillende niveaus (conditioneel op de overige predictoren in het model).

Merk op dat wanneer $I = 2$ er slechts 1 parameter getoetst moet worden. In dat geval is de F-toets ook equivalent aan de t-toets zoals in voorgaande sectie.

We hernemen het voorbeeld van de pijneducatie (sectie 1.2.2) waarbij we het effect van `conditie` (3 niveaus) op de uitkomst `buig` regresseren. De output bij dummy-codering is als volgt:

```
> fit_pijneducatie_dummy<-lm(Buig~Conditie,data=pijneducatie)
> summary(fit_pijneducatie_dummy)
```

```
Call:
lm(formula = Buig ~ Conditie, data = pijneducatie)

Residuals:
Min       1Q   Median       3Q      Max
-25.3164  -5.5974  -0.0914   4.7621  26.7199

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -0.7655     1.5885  -0.482   0.6308
ConditieAlgemene pijneducatie   3.3736     2.2464   1.502   0.1358
ConditieRugpijneducatie         5.8291     2.2327   2.611   0.0102 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.05 on 118 degrees of freedom
Multiple R-squared:  0.05497,Adjusted R-squared:  0.03895
F-statistic: 3.432 on 2 and 118 DF,  p-value: 0.03559
```

Aangezien `conditie` de enige predictor in het model is, lezen we hier onmiddellijk het resultaat voor de modelvergelijkingstoets voor het effect van `conditie` af (door vergelijking van het model met nulmodel): de geobserveerde toetsingsgrootheid f^* is gelijk aan 3.432 met een bijhorende p-waarde gelijk aan 0.036. We kunnen bijgevolg de nulhypothese die stelt dat de gemiddelde verschilscore in voorover buigen gelijk is in de 3 condities, verwerpen.

Dit resultaat kunnen we ook als volgt bekomen:

```
> fit0_pijneducatie_dummy<-lm(Buig~1,data=pijneducatie)
> anova(fit0_pijneducatie_dummy,fit_pijneducatie_dummy)
Analysis of Variance Table

Model 1: Buig ~ 1
Model 2: Buig ~ Conditie
Res.Df   RSS     Df Sum of Sq   F  Pr(>F)
1    120 12602
2    118 11910  2    692.76 3.4319 0.03559 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Kijk zelf na dat het resultaat hetzelfde blijft wanneer effect-codering gebruikt wordt voor

conditie.

4.4.3 Algemene strategie: Anova-tabel in R

Van zodra er meerdere predictoren (waaronder nominale predictoren) in een regressiemodel opgenomen zijn, is het praktischer om de resultaten voor de toetsen van alle predictoren afzonderlijk via 1 enkel commando te verkrijgen. Het commando `Anova` (let op de hoofdletter!) uit het R-package `car` geeft deze resultaten in 1 anova-tabel weer. De reden dat we werken met dit commando is dat het toelaat verschillende types toetsen uit voeren waaronder deze die overeenkomen met wat andere software zoals SPSS standaard zou weergeven.

Wij maken altijd gebruik van Type III-toetsen bij het toetsen van de predictoren afzonderlijk. Het onderscheid tussen de verschillende types toetsen speelt voornamelijk een rol wanneer ook interacties in het model opgenomen zijn. We komen er op terug in sectie 5 over interacties.

Een belangrijke opmerking bij het gebruik van de Type III-toetsen bij Anova is dat we voor het toetsen zelf eerst moeten overgaan op effect-codering van de nominale variabelen in het model. Onthoud dat een model met effect-codering in essentie hetzelfde is als een model met dummy-codering; enkel de interpretatie van de parameters wijzigt. Het betreft dus louter een technisch aspect om correcte en zinvolle resultaten bij het toetsen te krijgen. Dit is opnieuw enkel van belang als er ook interacties met nominale predictoren opgenomen zijn (zie verder).

We hernemen het voorbeeld rond pijneducatie (sectie 1.2.2) . In deze studie zijn de participanten niet at random toegewezen aan de verschillende condities. Daarom is het zinvol om te corrigeren voor het effect van leeftijd en de graad van depressie. Leeftijd (`Leeft`) en Depressiescore (`Dep`) mogen als van intervalniveau verondersteld worden.

```
> fit3_pijneducatie_dummy<-lm(Buig~Conditie+Leeft+Dep,data=pijneducatie)
> summary(fit3_pijneducatie_dummy)

Call:
lm(formula = Buig ~ Conditie + Leeft + Dep, data = pijneducatie)

Residuals:
Min       1Q   Median       3Q      Max
-20.2287  -4.9537  -0.5254   5.1798  18.7101

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 16.42433    4.41878   3.717 0.000312 ***
```

```
ConditieAlgemene pijneducatie  3.73261    1.72868   2.159 0.032891 *
ConditieRugpijneducatie        4.78874    1.72026   2.784 0.006276 **
Leeft                         -0.10238    0.10749  -0.952 0.342840
Dep                           -0.65121    0.07187  -9.061 3.75e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.723 on 116 degrees of freedom
Multiple R-squared:  0.451,Adjusted R-squared:  0.4321
F-statistic: 23.82 on 4 and 116 DF,  p-value: 2.133e-14
```

Uit bovenstaande output kunnen we aflezen dat leeftijd in het model geen statistisch significante invloed heeft op de uitkomst ($p = 0.34$) en depressiescore wel ($p < 0.001$), maar voor conditie kunnen we dit niet rechtstreeks aflezen. Wat we aflezen is het resultaat van de toetsen waarbij de conditie met algemene pijneducatie en rugpijneducatie afzonderlijk met de baseline conditie vergeleken worden.

We vragen nu de bijhorende anova-tabel op om de resultaten voor alle predictoren afzonderlijk te zien. Hiervoor herdefiniëren we eerst het model aan de hand van effect-codering.

```
> library(car)
> fit3_pijneducatie_test<-lm(Buig~Conditie+Leeft+Dep,data=pijneducatie,
                                  contrasts=list(Conditie=contr.sum))
> Anova(fit3_pijneducatie_test,type=3)
Anova Table (Type III tests)

Response: Buig
            Sum Sq  Df F value    Pr(>F)
(Intercept) 1208.3   1 20.2587 1.618e-05 ***
Conditie     509.2   2  4.2685   0.01626 *
Leeft         54.1   1  0.9072   0.34284
Dep         4897.2   1 82.1057 3.746e-15 ***
Residuals   6918.8 116
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

In deze output lezen we voor iedere predictor het resultaat af van de modelvergelijkingstoets waarbij de modellen met en zonder de predictor vergeleken worden (de geobserveerde toetsingsgrootheid in `F value` en p-waarde in `Pr(>F)`). We krijgen de overeenkomstige

kwadratensom voor deze toetsen (`Sum Sq`) en het aantal vrijheidsgraden (`Df`), i.e. het aantal parameters dat getoetst wordt (1 voor leeftijd en 1 voor depressiescore maar 2 voor conditie).

Voor `Leeft` en `Dep` bekomen we (uiteraard) dezelfde resultaten voor de F-toets als voor de t-toets. We lezen verder af dat het effect van conditie na correctie voor leeftijd en depressiescore nog steeds statistisch significant is op het 5% significantieniveau ($p = 0.016$).

Verifieer zelf dat via `Anova` dezelfde resultaten bekomen worden voor de analyses uitgevoerd in sectie 4.4.1 en 4.4.2!

5 Interactie (moderatie)

5.1 Wat is interactie?

Bij een regressiemodel met 2 predictoren X_1 en X_2 waarvoor

$$E(Y_i|x_{i1}, x_{i2}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ \ (i = 1, \ldots, n)$$

zijn de effecten van de predictoren additief: het effect van de ene predictor hangt niet af van het niveau van de andere, hun gezamenlijk effect kan gemodelleerd worden als een som.

Wanneer er een **interactie** is tussen beide predictoren betekent dit dat het effect van een combinatie van de 2 predictoren groter of kleiner is dan de som van de afzonderlijke effecten. Het effect van de ene predictor is nu anders voor elk niveau van de andere predictor. We hebben dan:

$$E(Y_i|x_{i1}, x_{i2}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} \ \ (i = 1, \ldots, n)$$

De interactieterm is gelijk aan het product van beide predictoren. β_3 is het interactie-effect.

De termen **effect-modificatie** en **moderatie** worden ook vaak gebruikt om te verwijzen naar een interactie. X_1 modificeert het effect van X_2 op Y (en vice versa). X_1 is een moderator voor het effect van X_2 op Y (en vice versa).

De regressiecoëfficiënten β_1 en β_2 hebben nu een andere interpretatie dan voorheen. Wanneer X_1 1 eenheid stijgt terwijl $X_2 = x_2$ constant blijft, neemt de verwachte uitkomst toe met $\beta_1 + \beta_3 x_2$.

Analoog: wanneer X_2 1 eenheid stijgt terwijl $X_1 = x_1$ constant blijft, neemt de verwachte uitkomst toe met $\beta_2 + \beta_3 x_1$.

Voorbeeld

Beschouwen we het verwachte aantal telefonisch verkochte abonnementen per dag in functie van de ervaring van de verkoper (X_1, gemeten op een schaal van 1 tot 7) en het instituut waarbij de verkoper een opleiding gekregen heeft (X_2, Instituut 1 en Instituut 2).

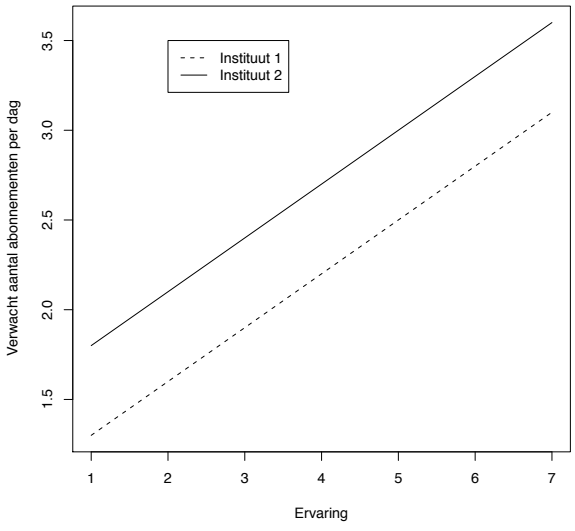
- Veronderstel dat $E(Y_i|x_{i1}, x_{i2}) = 1 + 0.3x_{i1} + 0.5x_{i2}$ waarbij $x_{i2} = 0$ voor instituut 1 en $x_{i2} = 1$ voor instituut 2 (dummy-codering). In dit geval is er geen interactie tussen de ervaring en het instituut waar de verkoper de opleiding gekregen heeft. We vinden immers dat het verwacht aantal verkochte abonnementen per dag in functie van de ervaring van een verkoper die een opleiding in instituut 1 kreeg als volgt is:

  $$E(Y_i|x_{i1}, x_{i2} = 0) = 1 + 0.3x_{i1}$$

  terwijl dit voor een verkoper die zijn opleiding in instituut 2 kreeg als volgt is:

  $$E(Y_i|x_{i1}, x_{i2} = 1) = 1 + 0.3x_{i1} + 0.5 = 1.5 + 0.3x_{i1}.$$

  Het effect van de ervaring blijft gelijk (0.3), alleen het intercept verandert. Grafisch kunnen we het verwacht aantal abonnementen per dag in functie van de ervaring als volgt weergeven:



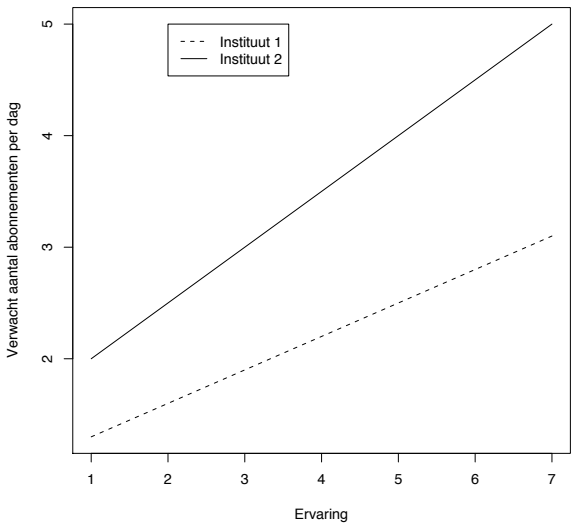
- In het geval dat $E(Y_i|x_{i1}, x_{i2}) = 1 + 0.3x_{i1} + 0.5x_{i2} + 0.2x_{i1}x_{i2}$ waarbij $x_{i2} = 0$ voor instituut 1 en $x_{i2} = 1$ voor instituut 2, is er wel een interactie tussen beide predictoren. Het verwacht aantal abonnementen per dag in functie van de ervaring van de verkoper die een opleiding kreeg in instituut 1 is immers

$$E(Y_i|x_{i1}, x_{i2} = 0) = 1 + 0.3x_{i1}$$

terwijl dit voor een verkoper die een opleiding kreeg in instituut 2 als volgt is

$$E(Y_i|x_{i1}, x_{i2} = 1) = 1 + 0.3x_{i1} + 0.5 + 0.2x_{i1} = 1.5 + 0.5x_{i1}.$$

We zien dat het effect van de ervaring groter is voor een verkoper die een opleiding in instituut 2 kreeg. Dit wordt getoond op onderstaande figuur.



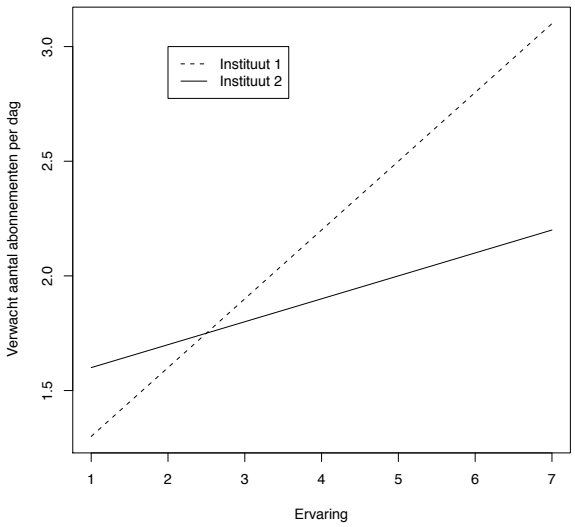
- In het geval dat $E(Y_i|x_{i1}, x_{i2}) = 1 + 0.3x_{i1} + 0.5x_{i2} - 0.2x_{i1}x_{i2}$ waarbij $x_{i2} = 0$ voor instituut 1 en $x_{i2} = 1$ voor instituut 2, is er ook een interactie tussen beide predictoren. Het verwacht aantal abonnementen per dag in functie van de ervaring van de verkoper die een opleiding kreeg in instituut 1 is dan

$$E(Y_i|x_{i1}, x_{i2} = 0) = 1 + 0.3x_{i1}$$

terwijl dit voor een verkoper die een opleiding kreeg in instituut 2 als volgt is

$$E(Y_i|x_{i1}, x_{i2} = 1) = 1 + 0.3x_{i1} + 0.5 - 0.2x_{i1} = 1.5 + 0.1x_{i1}.$$

Hier is het effect van de ervaring kleiner voor een verkoper die een opleiding in instituut 2 kreeg. Dit wordt getoond op onderstaande figuur.



Op de bovenstaande figuren zien we dat de regressierechten (voor verwacht aantal in functie van ervaring) voor instituut 1 en instituut 2 parallel zijn wanneer er geen interactie aanwezig is. Bij interactie zijn deze rechten niet langer parallel.

In de praktijk is het nuttig om een dergelijke plot te maken op basis van de geschatte regressievergelijkingen en/of geschatte groepsgemiddeldes. Op die manier kan onderzocht worden of er aanwijzingen zijn voor interacties (zie verder).

5.2 Hoofd- en interactie-effecten

In bovenstaand voorbeeld representeren β_1 en β_2 de *hoofdeffecten* van respectievelijk X_1 en X_2; β_3 stelt het *interactie-effect* voor. De aanwezigheid van een interactie impliceert dat effecten niet eenduidig te interpreteren zijn: het effect van X_1 hangt immers af van het niveau van X_2 en vice versa.

5.3 Implementatie en toetsen van interactie-effecten

Een term voor de interactie tussen 2 predictoren kan eenvoudig in een model toegevoegd worden door een nieuwe variabele te creëren waarvan ieder element het product is van de overeenkomstige elementen van de predictoren en die variabele in het model te stoppen.

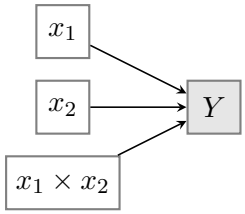
Het implementeren van interactie-effecten is routine en is vrij eenvoudig in statistische software zoals R.

Wanneer men echter interactietermen toevoegt aan het model, is het mogelijk dat er problemen met collineariteit ontstaan door de correlatie tussen de predictoren en de interactietermen. Collineariteit treedt op als twee (of meer) predictoren sterk met elkaar correleren. Dit wordt ook multicollineariteit genoemd en heeft als gevolg dat de standaardfouten van schatters heel groot kunnen worden. Dit euvel wordt gedeeltelijk verholpen door de predictoren van intervalniveau waarvoor een interactieterm aanwezig is eerst te centreren of nog: door de oorspronkelijke waarden van de predictoren $x_{i\ell}$ te vervangen door hun deviatiescore $x_{i\ell} - \bar{x}_\ell$. Dit zorgt er ook voor dat de parameters voor de hoofdeffecten makkelijker te interpreteren zijn.

Het toetsen van interactie-effecten gebeurt aan de hand van modelvergelijkingen zoals beschreven in de voorgaande sectie. In de aanwezigheid van interacties, maakt het wel een verschil uit welk type toets we gebruiken. De volgende types bestaan:

- **Type III toetsen**: effecten worden getoetst terwijl gecorrigeerd wordt voor alle andere effecten; i.e. het model met een effect wordt vergeleken met het model zonder dat effect.

  Concreet: stel dat in een regressiemodel met 2 predictoren een interactie tussen X_1 en X_2 opgenomen is, dan zal de toets voor het hoofdeffect van X_1 het volledige model (2 hoofdeffecten + interactie-effect) vergelijken met het model zonder het hoofdeffect van X_1 dus een model met een hoofdeffect van X_2 en het interactie-effect. De toets voor het interactie-effect vergelijkt het volledige model met het model zonder interactie-effect.

  Toetsing van hoofdeffect van X_2:

  - Via modelvergelijking:
    - model 1: $X_1 + X_2 + (X_1 \times X_2)$
    - model 2: $X_1 + (X_1 \times X_2)$

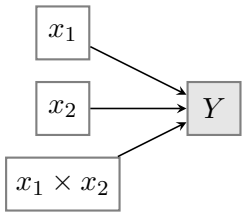
  Toetsing van interactie-effect:

  - Via modelvergelijking:
    - model 1: $X_1 + X_2 + (X_1 \times X_2)$

    * model 2: $X_1 + X_2$

- **Type II toetsen**: effecten worden getoetst terwijl gecorrigeerd wordt voor alle effecten van dezelfde of een lagere orde maar niet voor hogere orde effecten die het te toetsen effect omvatten.

  Concreet: stel dat in een regressiemodel met 2 predictoren een interactie tussen X_1 en X_2 opgenomen is, dan zal de toets voor het hoofdeffect van X_1 de interactie tussen X_1 en X_2 niet in rekening brengen aangezien dit een hogere orde term is die X_1 omvat. Hier wordt dan een model met X_1 en X_2 (zonder de interactieterm) vergeleken met een model met enkel X_2 (zonder de interactieterm). De toets voor het interactie-effect vergelijkt het volledige model (2 hoofdeffecten + interactie-effect) met het model zonder interactie-effect.



  Toetsing van hoofdeffect van X_2:

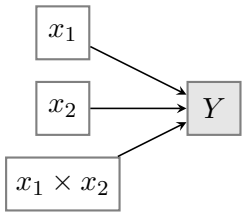
  – Via modelvergelijking:
    * model 1: $X_1 + X_2$
    * model 2: X_1

  Toetsing van interactie-effect:

  – Via modelvergelijking:
    * model 1: $X_1 + X_2 + (X_1 \times X_2)$
    * model 2: $X_1 + X_2$

- **Type I toetsen**: hier volgt het toetsen een sequentiële strategie en hangt dus af van de volgorde waarin predictoren aan een model toegevoegd worden.

  Concreet: als we eerst X_1 als predictor in een model voegen, dan X_2 en vervolgens het interactie-effect, zal de toets voor X_1 het model met enkel X_1 vergelijken met het nulmodel. De toets voor X_2 zal het model met X_1 en X_2 vergelijken met het model met enkel X_1 terwijl de toets voor het interactie-effect het volledige model (2 hoofdeffecten + interactie-effect) vergelijkt met het model zonder de interactie.



Toetsing van hoofdeffect van X_1:

- Via modelvergelijking:
  - model 1: X_1
  - model 2: nulmodel (geen predictoren)

Toetsing van hoofdeffect van X_2:

- Via modelvergelijking:
  - model 1: $X_1 + X_2$
  - model 2: X_1

Toetsing van interactie-effect:

- Via modelvergelijking:
  - model 1: $X_1 + X_2 + (X_1 \times X_2)$
  - model 2: $X_1 + X_2$

In R worden standaard de Type I toetsen gehanteerd. Daarom maken wij gebruik van `Anova` (package `car`) waarbij we kunnen aangeven of we Type II of Type III toetsen gebruiken. Uit de redenering hierboven kan afgeleid worden dat de toets voor het interactie-effect hetzelfde is maar er is een verschil tussen beide types voor het toetsen van de hoofdeffecten.

Wij gebruiken **Type III toetsen** aangezien dit de standaard is binnen andere softwarepakketten zoals SPSS.

Opmerking met betrekking tot nominale predictoren

Wanneer we interacties met nominale predictoren beschouwen, dienen ook extra hulpveranderlijken aangemaakt te worden.

In het algemeen, voor 2 nominale variabelen met respectievelijk I en J niveaus, hebben we $(I-1) \times (J-1)$ hulpveranderlijken nodig om het interactie-effect te representeren.

Om te onderzoeken of een interactie-effect statistisch significant is, moeten we een modelvergelijkingstoets uitvoeren om na te gaan of de regressiecoëfficiënten die horen bij de $(I-1) \times (J-1)$ hulpveranderlijken van de interactie 0 zijn.

Ongeacht of men dummy- of effect-codering gebruikt voor de nominale variabelen, geeft de toets voor het interactie-effect hetzelfde resultaat. Wanneer een interactieterm aanwezig is, zullen de toetsen voor de hoofdeffecten wel verschillen naargelang de codering die gehanteerd wordt. Bij dummy-codering wordt het effect van een nominale variabele getoetst binnen het referentieniveau van de andere nominale variabele. Bij effect-codering wordt het effect van een nominale variabele getoetst, uitgemiddeld over de niveaus van de andere nominale variabele.

Merk opnieuw op dat we in deze cursus toetsen aan de hand van **Type III toetsen** en dat we hierbij voor het toetsen overgaan op **effect-codering** van de nominale predictoren.

Opmerking met betrekking tot een combinatie van nominale predictoren en predictoren van intervalniveau

Als we een interactie tussen een nominale predictor van I niveaus en een predictor van intervalniveau beschouwen, zijn er $(I-1)$ hulpveranderlijken die dit interactie-effect representeren. Nagaan of het interactie-effect statistisch significant is, houdt dus in dat we een modelvergelijkingstoets uitvoeren om na te gaan of de regressiecoëfficiënten die horen bij de $(I-1)$ hulpveranderlijken van de interactie 0 zijn.

De hierboven beschreven situatie waarbij een verschil bestaat tussen toetsen voor hoofdeffecten wanneer men dummy- of effect-codering gebruikt, doet zich ook voor wanneer de interactie tussen een nominale predictor en een predictor van intervalniveau onderzocht wordt.

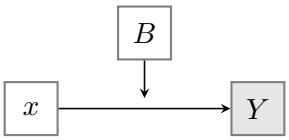
Opnieuw geldt hier dat we toetsen aan de hand van **Type III toetsen** en dat we hierbij voor het toetsen overgaan op **effect-codering** van de nominale predictoren.

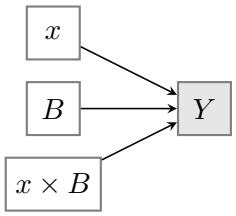
Verschillende gevallen

1. Interactie tussen 2 nominale predictoren (factoren) A en B

- toetsing: via modelvergelijking:
  - model 1: $A + B + (A \times B)$
  - model 2: $A + B$
- indien interactie $(A \times B)$ significant: moderatie

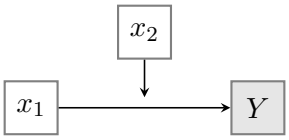
2. Interactie tussen een nominale predictor B en een predictor van intervalniveau x

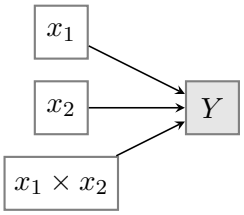




- toetsing: via modelvergelijking:
  - model 1: $x + B + (x \times B)$
  - model 2: $x + B$
- indien interactie $(x \times B)$ significant: moderatie

3. Interactie tussen 2 predictoren van intervalniveau x_1 en x_2





- toetsing: via modelvergelijking:
  - model 1: $x_1 + x_2 + (x_1 \times x_2)$
  - model 2: $x_1 + x_2$
- indien interactie $(x_1 \times x_2)$ significant: moderatie

5.4 Voorbeeld: herstel na coma

We illustreren de interpretatie en het toetsen van interactie-effecten in de verschillende gevallen aan de hand van het voorbeeld over het herstel na een coma (zie sectie 1.2.3).

We modelleren het wiskundig IQ (`piq`) in functie van verbaal IQ (`viq`), duur van de coma in dagen (hier gebruiken we de versie waarbij de 4 intervallen onderscheiden worden, `duration_cat`), gender van de patiënt (`sex`), leeftijd van de patiënt (`age`) .

`viq` en `age` zijn predictoren van intervalniveau, `sex` is een predictor van nominaal niveau en `duration_cat` beschouwen we ook als een predictor van nominaal niveau.

```
> class(coma$duration_cat)
[1] "factor"
> contrasts(coma$duration_cat)
          (1,7] (7,14] (14,255]
[0,1]         0      0        0
(1,7]         1      0        0
(7,14]        0      1        0
(14,255]      0      0        1

> class(coma$sex)
[1] "factor"
> contrasts(coma$sex)
       Male
Female    0
Male      1
```

Bij het interpreteren van de parameters zullen we dummy-codering gebruiken. Bij `duration_cat` is de kortste duur de referentiecategrie en bij `sex` is vrouw de referentiecategorie.

5.4.1 Het lineair regressiemodel zonder interacties

We bekijken eerst het lineair regressiemodel zonder interacties.

```
> fit1_coma<-lm(piq~viq+duration_cat+sex+age,data=coma)
> summary(fit1_coma)
```

```
Call:
lm(formula = piq ~ viq + duration_cat + sex + age, data = coma)

Residuals:
Min      1Q  Median      3Q     Max
-34.639  -5.847  -0.503   6.532  29.852

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          30.60636    6.41685   4.770 3.63e-06 ***
viq                   0.60328    0.06085   9.914  < 2e-16 ***
duration_cat(1,7]    -2.26030    2.15909  -1.047 0.296466
duration_cat(7,14]   -4.33226    2.28508  -1.896 0.059468 .
duration_cat(14,255] -8.28546    2.19158  -3.781 0.000209 ***
sexMale              -0.72454    1.91321  -0.379 0.705322
age                   0.03617    0.05705   0.634 0.526768
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.01 on 193 degrees of freedom
Multiple R-squared:  0.3962,Adjusted R-squared:  0.3774
F-statistic: 21.11 on 6 and 193 DF,  p-value: < 2.2e-16
```

De verschillende effecten in het model kunnen ook weergegeven worden aan de hand van de functie `effect` in het R package `effects`.

Voor `viq` zien we een positief effect ($\hat{b} = 0.60$). Dit betekent dat wanneer alle overige predictoren constant blijven, de gemiddelde `piq` toeneemt als `viq` toeneemt.

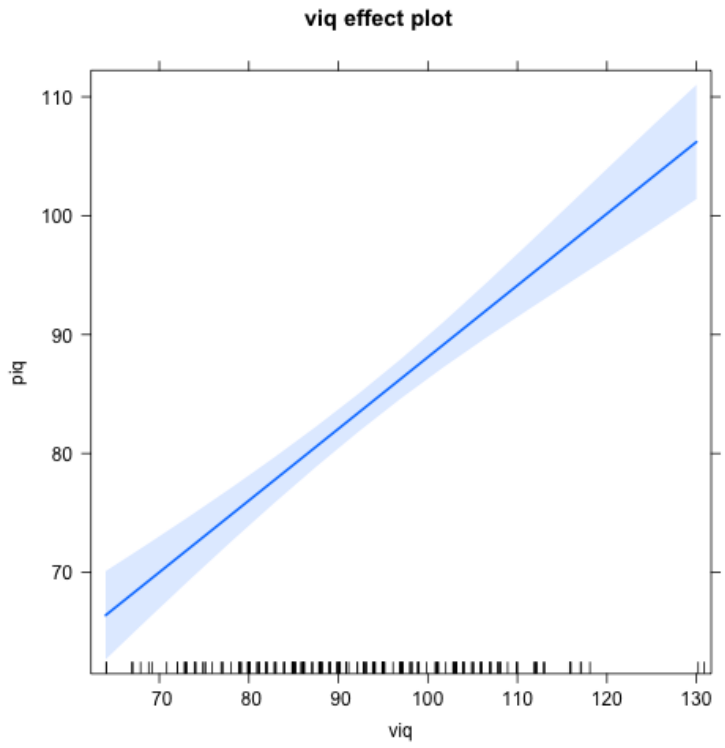
```
> library(effects)
> effect("viq",fit1_coma)

viq effect
viq
64        81        98       110       130
66.38895  76.64466  86.90037  94.13969 106.20524
```

Via `effects` wordt voor een aantal waarden van verbaal IQ (`viq` = 64, 81, 98, 110 , 130) het wiskundige IQ voorspeld: predictie `piq` = 66.39, 76.65, 86.90, 94.14, 106.21. Hierbij worden de

overige predictoren constant gehouden. Bij de predictie worden predictoren van intervalniveau gelijk gesteld aan hun gemiddelde waarde, terwijl een gewogen gemiddelde van de uitkomst berekend wordt over de verschillende groepen gevormd door de (combinaties van) nominale predictoren. Aangezien er geen interactieterm met `viq` in het model zit, maakt het niveau waarop de overige predictoren constant gehouden worden niet uit voor het effect van `viq`, enkel voor de voorspelde waarde van de uitkomst `piq`.

Het effect kan ook grafisch voorgesteld worden (met betrouwbaarheidsinterval):

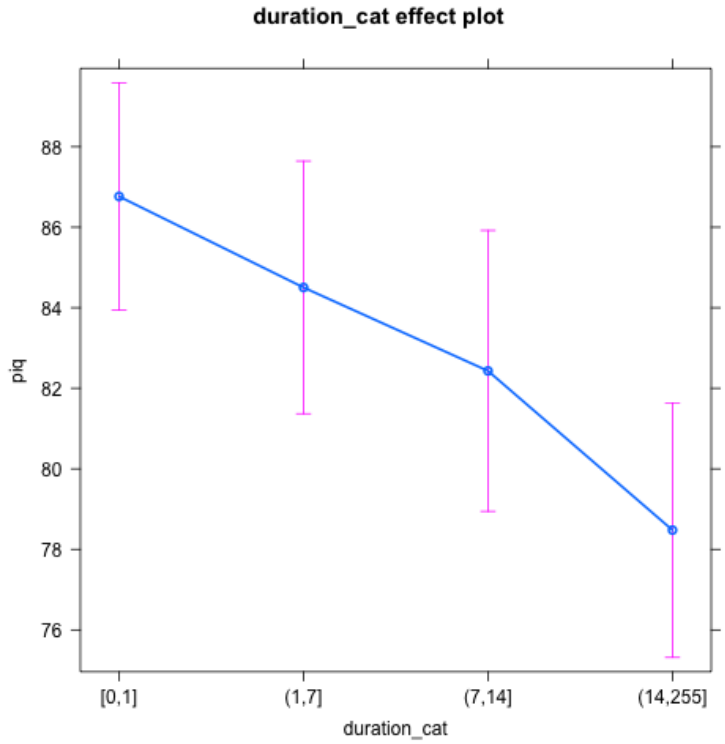
```
plot(effect("viq",fit1_coma))
```



Voor het effect van de duur van de coma (gecategoriseerd) krijgen we:

```
> effect("duration_cat",fit1_coma)

duration_cat effect
duration_cat
```

```
[0,1]    (1,7]    (7,14]  (14,255]
86.76350 84.50320 82.43124 78.47804
```



Hier zien we, net als op basis van de geschatte regressiecoëfficiënten, dat het geschatte gemiddelde wiskundig IQ afneemt naarmate de coma langer geduurd heeft.

5.4.2 Interactie tussen 2 nominale predictoren

We voegen de interactie tussen de duur van de coma (`duration_cat`) en gender (`sex`) in het model toe. Dit kunnen we doen door de term `duratio_cat:sex` in het model toe te voegen.

Opmerking: wanneer we `duration_cat*sex` *in het model toevoegen, zal R automatisch ook alle hoofdeffecten van de termen die in de interactie opgenomen zijn, toevoegen. Hier maakt dit geen verschil omdat deze hoofdeffecten al in het model zitten. Bovendien beschouwen wij geen modellen waar een interactieterm in zit zonder de hoofdeffecten van de termen die in de interactie opgenomen zijn.*

```
> fit2_coma<-lm(piq~viq+duration_cat+sex+age+duration_cat:sex,data=coma)
> summary(fit2_coma)

Call:
lm(formula = piq ~ viq + duration_cat + sex + age + duration_cat:sex,
data = coma)

Residuals:
Min      1Q  Median      3Q     Max
-33.802  -6.378  -0.374   5.962  27.361

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   35.26586    6.95253   5.072  9.3e-07 ***
viq                            0.58947    0.06127   9.620  < 2e-16 ***
duration_cat(1,7]             -7.29674    4.18644  -1.743  0.08296 .
duration_cat(7,14]            -6.49533    5.16375  -1.258  0.20998
duration_cat(14,255]         -15.80542    5.01042  -3.155  0.00187 **
sexMale                       -5.17123    3.44390  -1.502  0.13487
age                            0.04026    0.05717   0.704  0.48211
duration_cat(1,7]:sexMale      6.67411    4.84814   1.377  0.17025
duration_cat(7,14]:sexMale     2.76867    5.74587   0.482  0.63046
duration_cat(14,255]:sexMale   9.22301    5.50591   1.675  0.09556 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11 on 190 degrees of freedom
Multiple R-squared:  0.407,Adjusted R-squared:  0.3789
F-statistic: 14.49 on 9 and 190 DF,  p-value: < 2.2e-16
```

Om het interactie-effect te toetsen via `Anova` en Type III toetsen, moeten we zorgen dat bij de nominale predictoren effect-codering gebruikt wordt. Dit is louter om te toetsen, om het interactie-effect te interpreteren maken we gebruik van het bovenstaande geschatte model (dummy-codering).

```
> fit2_coma_test<-lm(piq~viq+duration_cat+sex+age+duration_cat:sex,
                  contrasts=list(duration_cat=contr.sum,sex=contr.sum),data=coma)
> Anova(fit2_coma_test,type=3)
Anova Table (Type III tests)
```

```
Response: piq
                  Sum Sq  Df F value    Pr(>F)
(Intercept)       2585.6   1 21.3623 7.005e-06 ***
viq              11202.0   1 92.5523 < 2.2e-16 ***
duration_cat      1947.3   3  5.3630  0.001446 **
sex                  7.9   1  0.0652  0.798744
age                 60.0   1  0.4960  0.482108
duration_cat:sex   417.7   3  1.1504  0.330054
Residuals        22996.6 190
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De toetsingsgrootheid voor het interactie-effect volgt onder de nulhypothese van geen interactie een F-verdeling met 3 en 190 vrijheidsgraden. De geobserveerde waarde is gelijk aan (417.7/3)/(22996.6/190)=1.15. De bijhorende p-waarde is gelijk aan 0.33.

Hoewel de interactie niet statistisch significant is op het 5% significantieniveau, gaan we ter illustratie na wat deze (niet-signifcante) interactie inhoudt.

1. **Geschatte effect van gender**: we bekijken op basis van het model het geschatte effect van gender voor verschillende categorieën van de duur van de coma. Een interactie-effect in het model impliceert dat dit geschatte effect zal wijzigen naargelang de duur van de coma.

   De overige predictoren (`viq`, `age`) die niet in het interactie-effect zitten, houden we constant en gelijk aan hun gemiddelde. Zoals de berekeningen hieronder tonen, hangt het geschatte effect van gender binnen een bepaalde duur niet af van het niveau waarop we deze variabelen constant houden aangezien er geen interactie met deze variabelen is.

   ```
   > mean(coma$viq)
   [1] 92.09
   > mean(coma$age)
   [1] 32.34668
   ```

   (a) Kortste duur van coma (`[0,1]`)

   - Voorspelde waarde uitkomst (`piq`) voor een vrouw in de categorie kortste duur:

   $$\hat{E}_1(Y) = 35.26586 + 0.58947 * 92.09$$
   $$\textit{intercept + verbaal IQ}$$

$$
\begin{aligned}
& -7.29674 * 0 - 6.49533 * 0 - 15.80542 * 0 \\
& \qquad \textit{kortste duur is referentie dus 3 hulpveranderlijken = 0} \\
& -5.17123 * 0 \\
& \qquad \textit{vrouw is referentie dus hulpveranderlijke = 0} \\
& +0.04026 * 32.34668 \\
& \qquad \textit{leeftijd} \\
& +6.67411 * 0 * 0 + 2.76867 * 0 * 0 + 9.22301 * 0 * 0 \\
& \qquad \textit{interactie tussen duur en gender} \\
= \; & 90.85243
\end{aligned}
$$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="[0,1]",
                          sex="Female", age=mean(coma$age))
> predict(fit2_coma,newdata)
90.85273
```

- Voorspelde waarde uitkomst (`piq`) voor een man in de categorie kortste duur:

$$
\begin{aligned}
\hat{E}_2(Y) \; = \; & 35.26586 + 0.58947 * 92.09 \\
& \qquad \textit{intercept + verbaal IQ} \\
& -7.29674 * 0 - 6.49533 * 0 - 15.80542 * 0 \\
& \qquad \textit{kortste duur is referentie dus 3 hulpveranderlijken = 0} \\
& -5.17123 * 1 \\
& \qquad \textit{vrouw is referentie dus hulpveranderlijke voor man = 1} \\
& +0.04026 * 32.34668 \\
& \qquad \textit{leeftijd} \\
& +6.67411 * 0 * 1 + 2.76867 * 0 * 1 + 9.22301 * 0 * 1 \\
& \qquad \textit{interactie tussen duur en gender} \\
= \; & 85.6812
\end{aligned}
$$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="[0,1]",
                        sex="Male",age=mean(coma$age))
> predict(fit2_coma,newdata)
1
85.6815
```

Effect van gender (mannen versus vrouwen) bij kortste duur:
85.6815-90.85273=-5.17123. Dit kunnen we ook rechtstreeks aflezen uit de geschatte

regressiecoëfficiënten in de output van het gehanteerde regressiemodel. Het geschatte hoofdeffect van gender geeft bij dummy-codering het effect van gender weer binnen het referentieniveau van de duur van de coma.

(b) Langste duur van coma (`(14,255]`)

- Voorspelde waarde uitkomst (`piq`) voor een vrouw in de categorie langste duur:

$$\begin{aligned}\hat{E}_3(Y) &= 35.26586 + 0.58947 * 92.09 \\ &\qquad \textit{intercept + verbaal IQ} \\ &\quad -7.29674 * 0 - 6.49533 * 0 - 15.80542 * 1 \\ &\qquad \textit{langste duur dus 3e hulpveranderlijke=1, de rest 0} \\ &\quad -5.17123 * 0 \\ &\qquad \textit{vrouw is referentie dus hulpveranderlijke = 0} \\ &\quad +0.04026 * 32.34668 \\ &\qquad \textit{leeftijd} \\ &\quad +6.67411 * 0 * 0 + 2.76867 * 0 * 0 + 9.22301 * 1 * 0 \\ &\qquad \textit{interactie tussen duur en gender} \\ &= 75.0473\end{aligned}$$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="(14,255]",
                                  sex="Female",age=mean(coma$age))
> predict(fit2_coma,newdata)
1
75.04731
```

- Voorspelde waarde uitkomst (`piq`) voor een man in de categorie langste duur:

$$\begin{aligned}\hat{E}_4(Y) &= 35.26586 + 0.58947 * 92.09 \\ &\qquad \textit{intercept + verbaal IQ} \\ &\quad -7.29674 * 0 - 6.49533 * 0 - 15.80542 * 1 \\ &\qquad \textit{langste duur dus 3e hulpveranderlijke=1, de rest 0} \\ &\quad -5.17123 * 1 \\ &\qquad \textit{vrouw is referentie dus hulpveranderlijke voor man = 1} \\ &\quad +0.04026 * 32.34668 \\ &\qquad \textit{leeftijd} \\ &\quad +6.67411 * 0 * 1 + 2.76867 * 0 * 1 + 9.22301 * 1 * 1 \\ &\qquad \textit{interactie tussen duur en gender}\end{aligned}$$

$= 79.09879$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="(14,255]",
                              sex="Male",age=mean(coma$age))
> predict(fit2_coma,newdata)
1
79.09909
```

Effect van gender (mannen versus vrouwen) bij langste duur: 79.09909-75.04731 =4.05178. Dit kunnen we ook afleiden uit de geschatte regressiecoëfficiënten in de output van het gehanteerde regressiemodel: -5.171232 (hoofdeffect gender) +9.22301 (interactie-effect met duur = langste duur)=4.051778.

2. **Geschatte effect van de duur van de coma**: we bekijken op basis van het model het geschatte effect van de duur van de coma voor mannen en vrouwen apart. Een interactie-effect in het model impliceert dat dit geschatte effect zal wijzigen naargelang gender.

   De overige predictoren (`viq`, `age`) die niet in het interactie-effect zitten, houden we opnieuw constant en gelijk aan hun gemiddelde.

   (a) Vrouwen

   Het geschatte verschil in gemiddeld wiskundig IQ bij vrouwen voor de langste versus de kortste duur is $\hat{E}_3(Y) - \hat{E}_1(Y) = 75.04731 - 90.85273 = -15.80542$. Dit is op afronding na de geschatte regressiecoëfficiënt die we aflezen bij het hoofdeffect van de hoogste categorie (i.e. het geschatte effect van de hoogste versus de laagste categorie bij vrouwen).

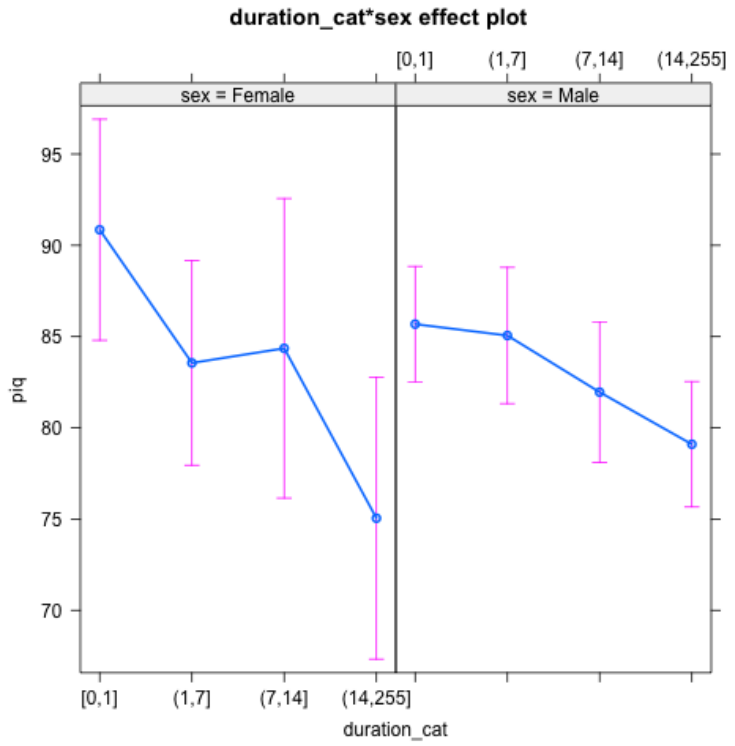
   (b) Mannen

   Het geschatte verschil in gemiddeld wiskundig IQ bij mannen voor de langste versus de kortste duur is $\hat{E}_4(Y) - \hat{E}_2(Y) = 79.09879 - 85.6812 = -6.58241$. Dit kunnen we (op afronding na) afleiden uit de geschatte regressiecoëfficiënten: -15.80542 (hoofdeffect hoogste categorie) + 9.22301 (interactie-effect met gender=man) = -6.58241.

In R krijgen we de volgende weergave van het interactie-effect:

```
> effect("duration_cat:sex", fit2_coma)

duration_cat*sex effect
```

```
sex
duration_cat   Female     Male
[0,1]         90.85273 85.68150
(1,7]         83.55599 85.05887
(7,14]        84.35740 81.95484
(14,255]      75.04731 79.09909
```



5.4.3 Interactie tussen een nominale predictor en een predictor van intervalniveau

Geval 1: de nominale predictor bestaat uit 2 niveaus

We voegen de interactie tussen gender (`sex`) en leeftijd (`age`) in het model toe.

```
> fit3_coma<-lm(piq~viq+duration_cat+sex+age+sex:age,data=coma)
> summary(fit3_coma)
```

```
Call:
lm(formula = piq ~ viq + duration_cat + sex + age + sex:age,
data = coma)

Residuals:
Min      1Q  Median      3Q     Max
-34.726  -5.869  -0.563   6.470  30.375

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          29.46295    6.94603   4.242 3.45e-05 ***
viq                   0.59959    0.06157   9.739  < 2e-16 ***
duration_cat(1,7]    -2.28337    2.16429  -1.055 0.292741
duration_cat(7,14]   -4.41770    2.29828  -1.922 0.056064 .
duration_cat(14,255] -8.35007    2.20121  -3.793 0.000199 ***
sexMale               1.20384    4.82652   0.249 0.803300
age                   0.08334    0.12249   0.680 0.497094
sexMale:age          -0.05973    0.13719  -0.435 0.663789
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.04 on 192 degrees of freedom
Multiple R-squared:  0.3968,Adjusted R-squared:  0.3748
F-statistic: 18.04 on 7 and 192 DF,  p-value: < 2.2e-16
```

De resultaten van de toetsen van de verschillende effecten kunnen hieronder afgelezen worden. Opnieuw: om het interactie-effect te interpreteren maken we gebruik van het bovenstaande geschatte model.

```
> fit3_coma_test<-lm(piq~viq+duration_cat+sex+age+age:sex,
                   contrasts=list(duration_cat=contr.sum,sex=contr.sum),data=coma)
> Anova((fit3_coma_test),type=3)
Anova Table (Type III tests)

Response: piq
              Sum Sq  Df F value    Pr(>F)
(Intercept)   2359.0   1 19.3629 1.791e-05 ***
viq          11554.9   1 94.8446 < 2.2e-16 ***
```

```
duration_cat  1863.0   3  5.0974  0.002042 **
sex              7.6   1  0.0622  0.803300
age             71.9   1  0.5899  0.443394
sex:age         23.1   1  0.1895  0.663789
Residuals    23391.2 192
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De toetsingsgrootheid voor het interactie-effect volgt onder de nulhypothese van geen interactie een F-verdeling met 1 en 192 vrijheidsgraden. De geobserveerde waarde is gelijk aan (23.1)/(23391/192)=0.19. De bijhorende p-waarde is gelijk aan 0.66.

Hoewel de interactie niet statistisch significant is op het 5% significantieniveau, gaan we ter illustratie opnieuw na wat deze (niet-signifcante) interactie inhoudt.

We leiden af wat het geschatte effect is van leeftijd op het gemiddelde wiskundig IQ. Een interactie-effect tussen gender en leeftijd impliceert dat dit geschatte effect anders zal zijn bij vrouwen en mannen.

1. **Geschatte effect van leeftijd bij vrouwen**: bij het afleiden van dit effect worden de predictoren die niet in de interactie zitten constant gehouden. Verbaal IQ (`viq`) stellen we gelijk aan het gemiddelde en we kijken binnen de laagste categorie van duur (`duration_cat`). Het geschatte effect van leeftijd hangt niet af van het niveau waarop we deze variabelen constant houden aangezien er geen interactie met deze variabelen is.

   ```
   > mean(coma$viq)
   [1] 92.09
   > mean(coma$age)
   [1] 32.34668
   ```

   (a) Voorspelde waarde uitkomst (`piq`) voor een vrouw met gemiddelde leeftijd in de categorie kortste duur:

   $$\begin{aligned}\hat{E}_{\overline{\text{age}}}(Y) &= 29.46295 + 0.59959 * 92.09 \\ &\qquad \textit{intercept + verbaal IQ} \\ &\quad -2.28337 * 0 - 4.41770 * 0 - 8.35007 * 0 \\ &\qquad \textit{kortste duur is referentie dus 3 hulpveranderlijken = 0} \\ &\quad +1.20384 * 0 \\ &\qquad \textit{vrouw is referentie dus hulpveranderlijke = 0}\end{aligned}$$

$$
\begin{aligned}
& +0.08334 * 32.34668 \\
& \qquad \textit{leeftijd} \\
& -0.05973 * 0 * 32.34668 \\
& \qquad \textit{interactie tussen gender en leeftijd} \\
= \; & 87.37497
\end{aligned}
$$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="[0,1]",
            sex="Female",age=mean(coma$age))
> predict(fit3_coma,newdata)
1
87.37468
```

(b) Voorspelde waarde uitkomst (`piq`) voor een vrouw in de categorie kortste duur wanneer leeftijd met 1 eenheid toeneemt:

$$
\begin{aligned}
\hat{E}_{\overline{\text{age}}+1}(Y) \; = \; & 29.46295 + 0.59959 * 92.09 \\
& \qquad \textit{intercept + verbaal IQ} \\
& +0.08334 * (32.34668 + 1) \\
& \qquad \textit{leeftijd} \\
= \; & 87.45831
\end{aligned}
$$

Via R:

```
> newdata<-data.frame(viq=mean(coma$viq),duration_cat="[0,1]",
             sex="Female",age=(mean(coma$age)+1))
> predict(fit3_coma,newdata)
1
87.45802
```

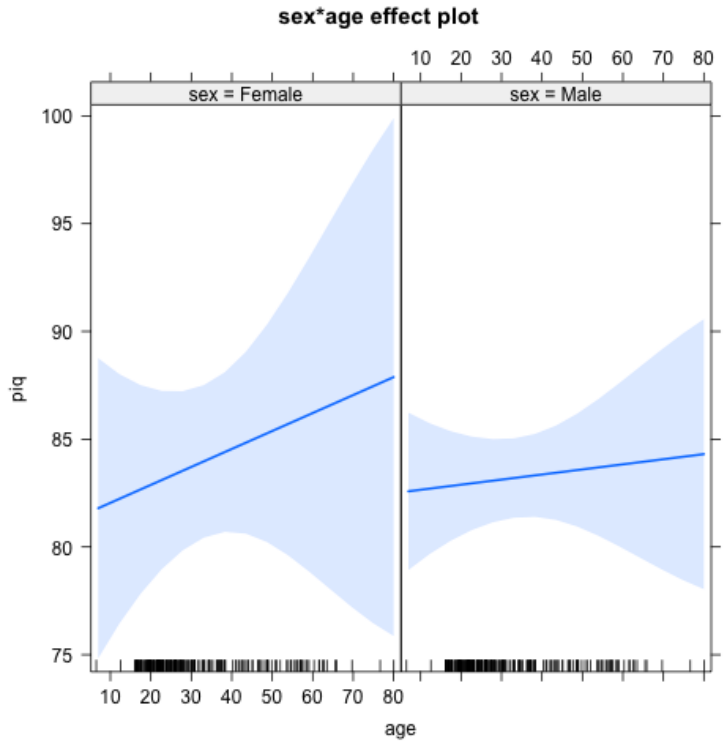
Het geschatte effect van leeftijd is gelijk aan 87.45802-87.37468=0.08334. Dit is de regressiecoëfficiënt die we aflezen bij het hoofdeffect van leeftijd (i.e. geschatte effect van leeftijd binnen referentieniveau van gender).

2. **Geschatte effect van leeftijd bij mannen**: opnieuw houden we verbaal IQ (`viq`) constant en gelijk aan het gemiddelde en we kijken binnen de laagste categorie van duur (`duration_cat`).

   Analoog aan voorgaande berekeningen kunnen we afleiden dat het geschatte effect gelijk is aan 0.08334 (hoofdeffect van leeftijd)-0.05973 (interactie-effect met gender=man)=0.02361. Dit geschatte effect van leeftijd bij mannen is kleiner dan dit effect bij vrouwen.

In R krijgen we de volgende weergave van het interactie-effect:

```
> effect("sex:age",fit3_coma)

sex*age effect
         age
sex         7       20       40       60       80
Female 81.79573 82.87911 84.54584 86.21257 87.87930
Male   82.58148 82.88839 83.36056 83.83273 84.30489
```



Geval 2: de nominale predictor bestaat uit meer dan 2 niveaus

We voegen de interactie tussen de duur van de coma (`duration_cat`) en leeftijd (`age`) in het model toe.

```
> fit4_coma<-lm(piq~viq+duration_cat+sex+age+duration_cat:age,data=coma)
```

```
> summary(fit4_coma)

Call:
lm(formula = piq ~ viq + duration_cat + sex + age + duration_cat:age,
data = coma)

Residuals:
Min      1Q  Median      3Q     Max
-34.907  -6.010  -0.659   6.805  29.498

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               31.70527    6.64696   4.770 3.66e-06 ***
viq                        0.60052    0.06175   9.725  < 2e-16 ***
duration_cat(1,7]         -2.02242    5.68100  -0.356   0.7222
duration_cat(7,14]        -0.39656    6.18241  -0.064   0.9489
duration_cat(14,255]     -13.33837    5.11249  -2.609   0.0098 **
sexMale                   -0.79532    1.91966  -0.414   0.6791
age                        0.01567    0.08890   0.176   0.8603
duration_cat(1,7]:age     -0.01379    0.16101  -0.086   0.9319
duration_cat(7,14]:age    -0.12992    0.17290  -0.751   0.4533
duration_cat(14,255]:age   0.17453    0.14568   1.198   0.2324
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.02 on 190 degrees of freedom
Multiple R-squared:  0.4055,Adjusted R-squared:  0.3774
F-statistic:  14.4 on 9 and 190 DF,  p-value: < 2.2e-16
```

Op basis van bovenstaande output zien we dat het geschatte effect van leeftijd (voor een constant verbaal IQ en gender) binnen de kortste duur (0 tot 1 dag) gelijk is aan 0.01567, binnen de categorie met een duur van meer dan 1 dag tot 7 dagen 0.01567-0.01379=0.00188, binnen de categorie met een duur van meer dan 7 dagen tot 14 dagen 0.01567-0.12992=-0.11425 en binnen de categorie met een duur van meer dan 14 dagen 0.01567+0.17453=0.1902. Het opnemen van een interactie-effect tussen duur en leeftijd impliceert dat het geschatte effect van leeftijd varieert naargelang de duur van de coma.

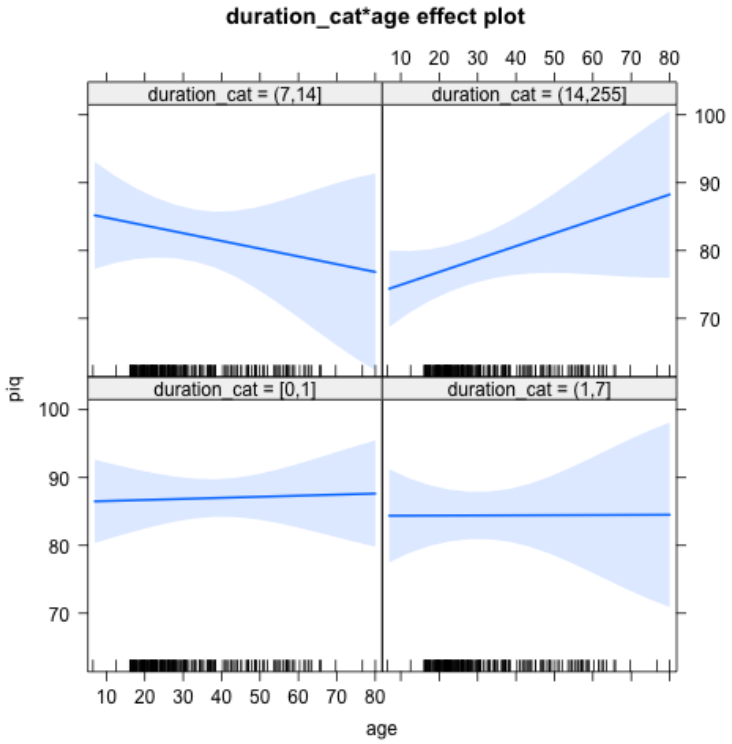
De resultaten voor de toetsen van de verschillende effecten in het model kunnen we aflezen in onderstaande output:

```
> fit4_coma_test<-lm(piq~viq+duration_cat+sex+age+duration_cat:age,
               contrasts=list(duration_cat=contr.sum,sex=contr.sum),data=coma)
> Anova((fit4_coma_test),type=3)
Anova Table (Type III tests)

Response: piq
                   Sum Sq  Df F value    Pr(>F)
(Intercept)       2406.5   1 19.8338 1.439e-05 ***
viq              11474.2   1 94.5691 < 2.2e-16 ***
duration_cat      1058.3   3  2.9074   0.03591 *
sex                 20.8   1  0.1716   0.67912
age                 17.6   1  0.1450   0.70378
duration_cat:age   361.4   3  0.9930   0.39728
Residuals        23052.9 190
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De toetsingsgrootheid voor het interactie-effect volgt onder de nulhypothese van geen interactie een F-verdeling met 3 en 190 vrijheidsgraden. De geobserveerde waarde is gelijk aan (361.4/3)/(23052.9/190)=0.99. De bijhorende p-waarde is gelijk aan 0.397.

De weergave van deze (niet-significante) interactie is als volgt:

```
> effect("duration_cat:age",fit4_coma)

duration_cat*age effect
                  age
duration_cat  7       20       40       60       80
[0,1]    86.49294 86.69667 87.01010 87.32353 87.63696
(1,7]    84.37403 84.39855 84.43628 84.47401 84.51174
(7,14]   85.18694 83.70172 81.41676 79.13181 76.84685
(14,255] 74.37626 76.84886 80.65285 84.45684 88.26083
```



5.4.4 Interactie tussen 2 predictoren van intervalniveau

We voegen de interactie tussen verbaal IQ `viq` en leeftijd `age` in het model toe.

```
> fit5_coma<-lm(piq~viq+duration_cat+sex+age+viq:age,data=coma)
> summary(fit5_coma)

Call:
lm(formula = piq ~ viq + duration_cat + sex + age + viq:age,
data = coma)

Residuals:
Min      1Q  Median      3Q     Max
-33.762  -5.993  -0.340   6.467  30.651

Coefficients:
```

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          38.121018  14.374328   2.652 0.008670 **
viq                   0.517736   0.158542   3.266 0.001294 **
duration_cat(1,7]    -2.071854   2.186682  -0.947 0.344581
duration_cat(7,14]   -4.094544   2.324836  -1.761 0.079793 .
duration_cat(14,255] -8.123493   2.212750  -3.671 0.000313 ***
sexMale              -0.580379   1.932286  -0.300 0.764228
age                  -0.197554   0.403956  -0.489 0.625366
viq:age               0.002572   0.004400   0.584 0.559586
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.03 on 192 degrees of freedom
Multiple R-squared:  0.3973,Adjusted R-squared:  0.3753
F-statistic: 18.08 on 7 and 192 DF,  p-value: < 2.2e-16
```

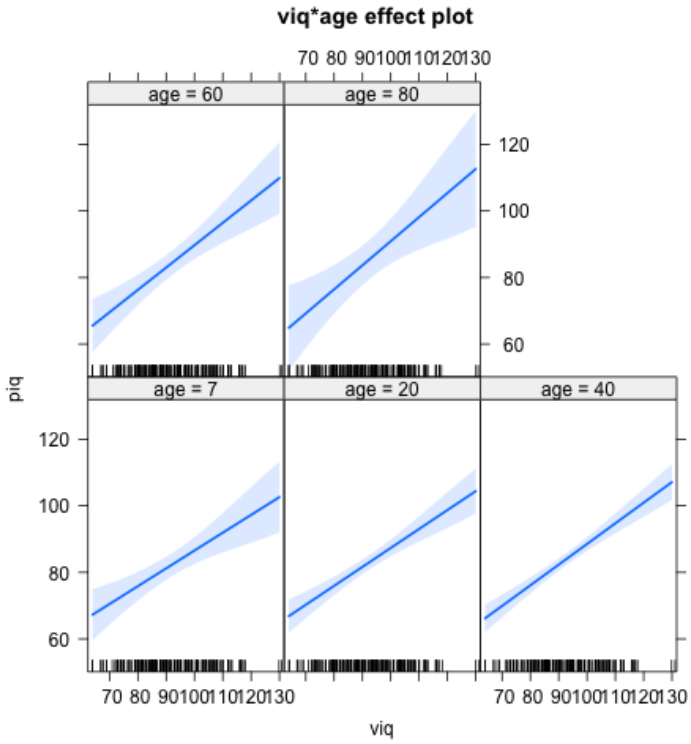
Het opnemen van het interactie-effect tussen verbaal IQ en leeftijd in het model impliceert dat het geschatte effect van verbaal IQ zal variëren naargelang de leeftijd en vice versa.

Analoog aan de afleidingen in sectie 5.1, zien we dat het geschatte effect van verbaal IQ gelijk is aan 0.517736+0.002572*`age`. Het geschatte effect van `viq` op het wiskundig IQ is bijgevolg positief en neemt lichtjes toe naarmate de leeftijd toeneemt.

Omgekeerd zien we ook dat het geschatte effect van leeftijd gelijk is aan -0.197554+0.002572*`viq`. Voor kleine waarden van `viq`, is het geschatte effect van leeftijd negatief; bij hoge waarden van `viq` zal dit positief worden.

Deze bevindingen zien we ook in de volgende output:

```
> effect("viq:age",fit5_coma)

viq*age effect
      age
viq        7        20        40        60        80
64   67.27360  66.84527  66.18629  65.52732  64.86835
81   76.38118  76.52125  76.73674  76.95223  77.16772
98   85.48876  86.19723  87.28718  88.37714  89.46710
110  91.91763  93.02733  94.73456  96.44178  98.14901
130 102.63243 104.41084 107.14684 109.88285 112.61886
```



Op de figuur zien we dat de rechten die het geschatte effect (slope) van `viq` weergeven voor verschillende leeftijd quasi parallel zijn. Dit komt overeen met een interactie-effect dat quasi 0 is. Op basis van de plots vinden we geen evidentie voor een bestaand interactie-effect tussen verbaal IQ en leeftijd m.b.t. hun effect op het gemiddeld wiskundig IQ.

Wanneer we het interactie-effect toetsen, zien we inderdaad dat dit niet statistisch significant is op het 5% significantieniveau ($p = 0.56$).

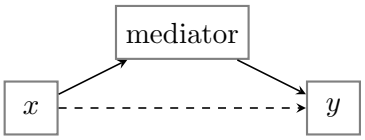
```
> fit5_coma_test<-lm(piq~viq+duration_cat+sex+age+viq:age,
            contrasts=list(duration_cat=contr.sum,sex=contr.sum),data=coma)
> Anova(fit5_coma_test,type=3)
Anova Table (Type III tests)

Response: piq
              Sum Sq  Df F value   Pr(>F)
(Intercept)    678.6   1  5.5748 0.019222 *
viq           1298.2   1 10.6642 0.001294 **
```

```
duration_cat  1758.0   3  4.8139 0.002960 **
sex             11.0   1  0.0902 0.764228
age             29.1   1  0.2392 0.625366
viq:age         41.6   1  0.3416 0.559586
Residuals    23372.7 192
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

6 Mediatie

6.1 Wat is mediatie?

- Mediatie = interventie:
  - De relatie tussen y en x wordt *gemedieerd* door een derde variabele
  - of nog: de mediator *intervenieert* in de relatie tussen y en x.



- Mediatie impliceert een conceptuele causale hypothese (*the mediational hypothesis*): de onafhankelijke variabele x beïnvloedt de mediator, en de mediator beïnvloedt de afhankelijke variabele y.
- Mediatie probeert te verklaren waarom x een invloed heeft op y.
- Mediatie-effecten zijn alomtegenwoordig in de gedragswetenschappen!
- In wat volgt beschouwen we zowel y als m (de mediator) als variabelen van minstens intervalniveau.

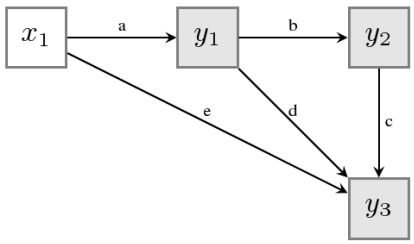
Een 'klassieke' paper waarbij het onderscheid tussen mediatie en moderatie aan bod komt, is:

> Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1185.

- Deze paper is al enorm veel geciteerd (zie Web of Science).
- De paper geeft een uitstekende beschrijving van het begrip *moderatie* en *mediatie*.
- Webpagina's:
  - `http://davidakenny.net/cm/moderation.htm`
  - `http://davidakenny.net/cm/mediate.htm`
- De auteurs beschrijven ook een 'statistische procedure' om na te gaan of er inderdaad sprake is van mediatie. Deze aanpak is bijzonder populair aangezien ze makkelijk uitvoerbaar is.
- In dit stuk nemen we de notatie van de auteurs over.

Paddiagrammen

- De schatting van een direct of rechtstreeks effect is een **padcoëfficiënt** (analoog met regressiecoëfficiënten)



- e representeert het direct effect van x_1 op y_3.
- Het indirect effect van x_1 op y_3:

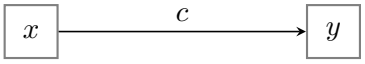
$$(a \times b \times c) + (a \times d)$$

- Het totaal effect van x_1 op y_3: direct + indirect:

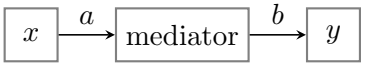
$$e + (a \times b \times c) + (a \times d)$$

Drie mogelijkheden m.b.t. mediatie

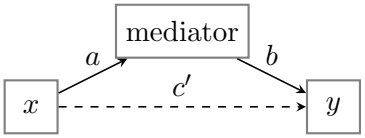
1. Geen mediatie: c = totaal effect van x op y:



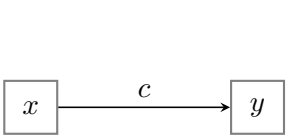
2. Volledige mediatie: $a \times b$ is het gemedieerd effect van x op y

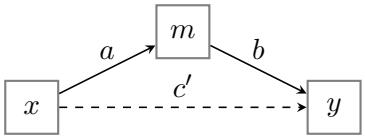


3. Gedeeltelijke mediatie: *totaal* effect: $c' + a \times b$, *direct* effect: c'



6.2 De Baron & Kenny methode





1. Regresseer y op x: is er wel een verband tussen y en x?

   1. $H_0 : c = 0$?

2. Regresseer m op x: is er wel een verband tussen m en x?

   2. $H_0 : a = 0$?

3. Regresseer y op x en m: is er wel een effect van m op y na controle voor x?

   3. $H_0 : b = 0$?

4. Is er na controle van m nog wel een effect van x op y?

   4. $H_0 : c' = 0$?

- Er is sprake van volledige mediatie indien:
  1. $a \neq 0$, $b \neq 0$ en $c \neq 0$,
  2. $c' = 0$
- Er is sprake van gedeeltelijke mediatie indien:
  1. $a \neq 0$, $b \neq 0$ en $c \neq 0$,
  2. $c' < c$
- Bij deze mediatie-analyse maakt men de assumptie van lineaire relaties tussen de verschillende variabelen.
- Het verschil $(c - c')$ hanteert men vaak als een maat voor het *gemedieerd* effect.
- Voor lineaire modellen geldt: onder normale omstandigheden (geen missing values, zelfde covariaten in beide modellen)
$$(c - c') = a \times b.$$
- Stap 1 is eigenlijk overbodig: indien de mediator fungeert als een 'suppressor', dan zal er geen verband zijn tussen x en y, terwijl er wel sprake is van mediatie; een symptoom is dat het teken van $a \times b$ omgekeerd is aan c'

  Een 'suppressor' of onderdrukkende variabele onderdrukt of verbergt de samenhang tussen 2 variabelen zodat deze geen verband met elkaar lijken te hebben.

  Omwille van deze reden gebruiken we in deze cursus de Baron & Kenny methode als volgt: we toetsen het totale effect c in stap 1, maar ongeacht het resultaat gaan we over naar de volgende stappen, omdat er nog steeds sprake kan zijn van mediatie. We voeren dus altijd de 4 stappen uit en gaan op basis van de resultaten na of er (1) evidentie is voor mediatie (stappen 2 en 3) (2) indien wel, of er sprake is van gedeeltelijke of volledige mediatie of dat er sprake is van suppressie (stappen 1 en 4).

6.3 De Sobel test

Enkele problemen met de Baron & Kenny methode:

1. Conceptueel: het is een indirecte wijze om het mediatie effect na te gaan: we focussen op het verschil tussen c en c', terwijl we eigenlijk geïnteresseerd zijn in a en b.
2. Statistisch:

  - in vaak voorkomende situaties: weinig power (MacKinnon et al., 2002)
  - meerdere toetsen na elkaar, dus een inflatie van Type I fouten
  - mogelijkheid tot inconsistente resultaten over de verschillende regressies heen:
    (a) a en b kunnen beide significant zijn, maar c' is niet kleiner dan c.
    (b) c' is veel kleiner dan c, maar a en b zijn niet significant
  - We bekomen enkel evidentie voor mediatie maar geen zekerheid; er zijn mogelijks andere modellen mogelijk! (bvb. m en y wisselen van plaats)

- Meer directe test: is het zo dat x een effect heeft op m ($= a$) en m een effect heeft op y ($= b$); met andere woorden, is $a \times b$ verschillend van nul?

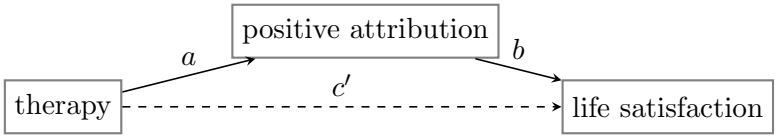
$$H_0 : a \times b = 0$$

  Deze test is de Sobel test.

- Probleem: de steekproevenverdeling van $(a \times b)$ is doorgaans niet normaal verdeeld (vooral bij kleinere steekproefgroottes).
- Via bootstrap (subsamplen uit een dataset) is het mogelijk om een 'empirische' steekproevenverdeling voor $(a \times b)$ te berekenen.
- MacKinnon et al. (2004) hebben een exacte verdeling afgeleid voor de steekproevenverdeling van $(a \times b)$.

6.4 Voorbeeld

We beschouwen de fictieve data uit de paper van Preacher & Hayes (2004) waarbij men geïnteresseerd is in het effect van een nieuwe cognitieve therapie (`therapy`, nominaal) op de levenstevredenheid (`satisfaction`, van intervalniveau) na pensionering (zie sectie 1.2.4). De onderzoeksvraag is of het effect van de cognitieve gedragstherapie gemedieerd wordt door de positiviteit van de attributies (`attribution`, van intervalniveau).



We voeren de 4 stappen van de Baron & Kenny methode uit om na te gaan of er al dan niet sprake is van mediatie.

- Stap 1

```
> satis_fitxy<-lm(satis~therapy,data=satisfaction)
> summary(satis_fitxy)

Call:
lm(formula = satis ~ therapy, data = satisfaction)

Residuals:
Min      1Q  Median      3Q     Max
-1.5669 -0.7319  0.3171  0.5121  1.3131

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3271     0.2233  -1.465   0.1541
therapy       0.7640     0.3058   2.498   0.0186 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8356 on 28 degrees of freedom
Multiple R-squared:  0.1823,Adjusted R-squared:  0.1531
F-statistic: 6.242 on 1 and 28 DF,  p-value: 0.01862
```

De p-waarde van de toets voor $H_0 : c = 0$ (totaal effect van therapie op satisfactie) is gelijk aan 0.019; we kunnen H_0 bijgevolg verwerpen op het 5% significantieniveau.

- Stap 2

```
> satis_fitxm<-lm(attrib~therapy,data=satisfaction)
> summary(satis_fitxm)

Call:
lm(formula = attrib ~ therapy, data = satisfaction)

Residuals:
Min      1Q  Median      3Q     Max
-1.4864 -0.5939 -0.0650  0.2611  1.8850

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  -0.3536     0.2184  -1.619   0.1166
therapy       0.8186     0.2990   2.738   0.0106 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8171 on 28 degrees of freedom
Multiple R-squared:  0.2111,Adjusted R-squared:  0.183
F-statistic: 7.494 on 1 and 28 DF,  p-value: 0.01064
```

De p-waarde voor de toets $H_0 : a = 0$ (effect van therapie op attributie) is gelijk aan 0.011; we kunnen H_0 bijgevolg verwerpen op het 5% significantieniveau.

- Stap 3 + 4

```
> satis_fitxmy<-lm(satis~therapy+attrib,data=satisfaction)
> summary(satis_fitxmy)

Call:
lm(formula = satis ~ therapy + attrib, data = satisfaction)

Residuals:
Min       1Q   Median       3Q      Max
-1.36527 -0.60758  0.02416  0.54923  1.29091

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1843     0.2185  -0.844    0.406
therapy       0.4334     0.3221   1.346    0.190
attrib        0.4039     0.1808   2.234    0.034 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7818 on 27 degrees of freedom
Multiple R-squared:  0.3098,Adjusted R-squared:  0.2587
F-statistic:  6.06 on 2 and 27 DF,  p-value: 0.006697
```

De p-waarde voor de toets $H_0 : b = 0$ (effect van attributie op satisfactie, na controle voor therapie) is gelijk aan 0.034; we kunnen H_0 bijgevolg verwerpen op het 5% significantieniveau. Op basis van de tot nu toe uitgevoerde stappen kunnen we besluiten dat er volgens de Baron & Kenny methode aanwijzingen voor mediatie zijn.

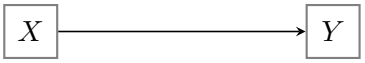
> Verder zien we dat, na correctie voor attributie, het effect van therapie op satisfactie niet statistisch significant is op het 5% significantieniveau (p-waarde gelijk aan 0.19). We hebben dus evidentie voor volledige mediatie.

We kunnen afleiden dat $\hat{c} = 0.76$ (i.e. het geschatte totaal effect van therapie op satisfactie) en $\hat{c}' = 0.43$ (i.e. het geschatte effect van therapie op satisfactie na correctie voor attributie). Het geschatte gemedieerde effect is bijgevolg $0.76 - 0.43 = 0.33$.

We bekijken in een volgende stap enkel de resultaten voor de Sobel test. In R krijgen we voor de bootstrap methode het volgende 95% betrouwbaarheidsinterval voor $a \times b$: $[0.081, 0.803]$ (we gaan niet dieper in op de manier waarop deze resultaten bekomen worden). Er is geen sprake van mediatie indien $a \times b = 0$. Aangezien het betrouwbaarheidsinterval 0 niet omvat, kunnen we $H_0 : a \times b = 0$ verwerpen op het 5% significantieniveau en aannemen dat er sprake is van mediatie.

7 De 'derde' variabele

Veronderstel dat we geïnteresseerd zijn in het effect van een predictor X op een uitkomst Y.



Een andere predictor Z (de 'derde' variabele) kan X en/of Y of de relatie tussen X en Y op verschillende manieren beïnvloeden. We bekijken hier enkele mogelijkheden, merk op dat Z een set van predictoren kan voorstellen.

7.1 Confounding

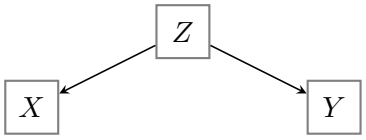
Regressie wordt vaak gebruikt om causale verbanden te onderzoeken, maar causale verklaringen zijn niet zomaar gerechtvaardigd.

Regressiemodellen beschrijven associaties die niet noodzakelijk interpreteerbaar zijn als causale effecten. Causale besluitvorming is mogelijk onder specifieke assumpties die niet getest kunnen worden a.d.h.v. data maar die soms gegarandeerd worden door het design (denk aan experimentele versus observationele designs).

Voorbeeld

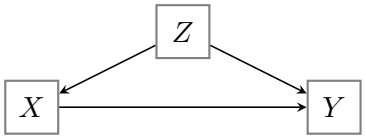
Via een regressie-analyse stelt men vast dat landen met een groter aantal televisies per persoon (X) een hogere levensverwachting (Y) hebben. Dit impliceert niet dat een hoger aantal televisies per persoon een hogere levensverwachting veroorzaakt. X is een indicator voor de welvaart. Welvaart beïnvloedt zowel X als Y. Wat men waarneemt is dus niet het rechtreekse verband tussen predictor en uitkomst.

In bovenstaand voorbeeld is er sprake van confounding.



Er is geen verband tussen X en Y, enkel een associatie omwille van de gemeenschappelijke oorzaak Z. Men spreekt van een spurieuze assocatie.

Confounding betekent niet noodzakelijk dat X geen effect heeft op Y, enkel dat minstens een deel van de associatie tussen X en Y verklaard wordt door het feit dat Z een confounder is voor de relatie tussen X en Y.



Om het effect van X op Y conditioneel op Z te schatten, moet Z als predictor in het regressiemodel opgenomen worden. Zie ook verder in sectie 7.4.

7.2 Moderatie

Het effect van X op Y hangt af van Z. Z is moderator voor de relatie tussen X en Y.

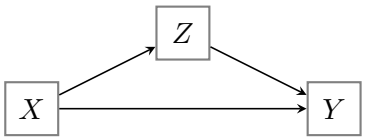
Voorbeeld

Cognitieve gedragstherapie is efficiënter bij adolescenten dan bij volwassenen.

Moderatie betekent dat er een interactie is tussen X en Z. De term 'interactie' is algemener aangezien de term 'moderatie' expliciet een onderscheid maakt tussen de rol van de variabelen. Hier: we zijn geïnteresseerd in het effect van X (predictor) op Y, maar deze relatie wijzigt naargelang het niveau van moderator Z. Zie sectie 5 voor het opnemen van interacties in een regressiemodel.

7.3 Mediatie

X is oorzaak van Z, Z is oorzaak van Y (Z is een *mediator* voor de relatie tussen X en Y)



Voorbeeld

Opleidingsniveau bepaalt motivatie en motivatie bepaalt leergierigheid.

Zie sectie 6 over het opsplitsen van het totaal effect van X op Y in een indirect of gemedieerd effect en een direct effect.

Merk op dat de data niet toelaten om de richting van de relaties na te gaan. Dit impliceert ook dat de data geen uitsluitsel kunnen bieden over het feit of Z een confounder of mediator is. De richting van de relaties kan gebaseerd zijn op onderliggende theorieën of gegarandeerd worden via het design via temporele ordening (X is gemeten op tijdstip 1, Z is gemeten op een later tijdstip 2 en Y is laatst gemeten op tijdstip 3).

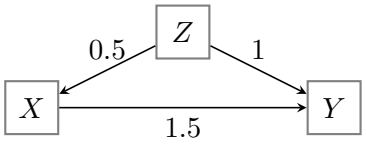
7.4 Omitted variable bias

De term 'omitted variable bias' verwijst naar de vertekening of bias bij het schatten van het effect van X op Y wanneer een variabele Z die geassocieerd is met zowel X als Y niet in het model opgenomen is.

Veronderstel bvb. dat Z een confounder is voor de relatie tussen X en Y en dat geldt voor $i = 1, \ldots, n$:

$$X_i \quad = \quad 1 + 0.5 \times Z_i + \varepsilon_i^* \text{ met } \varepsilon_i^* \sim N(0, 1)$$

$$Y_i \quad = \quad 1 + 1.5 \times X_i + Z_i + \varepsilon_i \text{ met } \varepsilon_i \sim N(0,1)$$

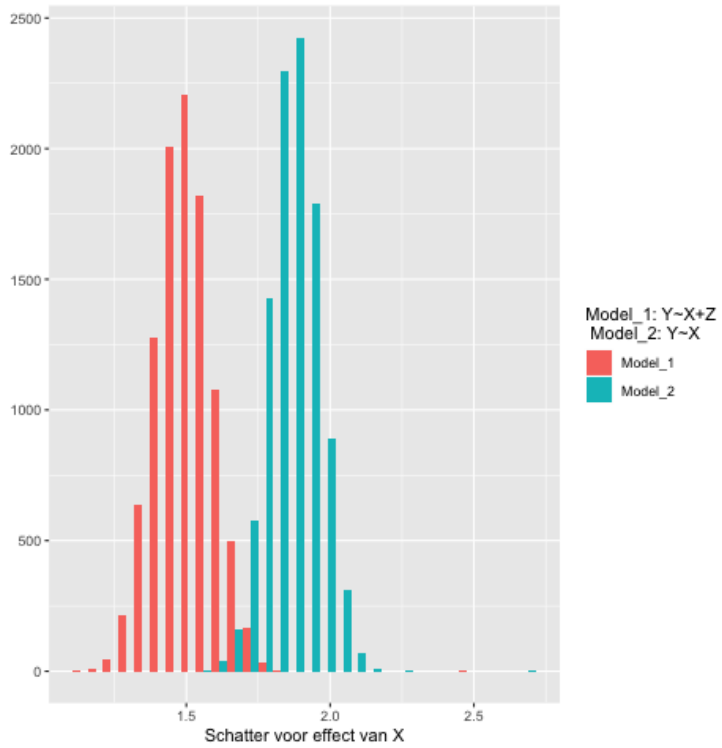


In dit voorbeeld kennen we de onderliggende waarheid, namelijk dat de regressieparameter die het effect van X op Y weergeeft gelijk is aan 1.5.

In een volgende stap genereren we 10 000 datasets waarbij X en Z vast zijn en gegenereerd volgens bovenstaande model en waarbij in iedere dataset de uitkomst Y volgens bovenstaand model gegenereerd is.

Voor iedere gegenereerde dataset schatten we 2 regressiemodellen. Model 1 is correct en bevat zowel X en Z als predictoren; model 2 bevat enkel X en is dus verkeerd gespecifieerd.

Onderstaande figuur toont voor beide modellen de steekproevenverdeling van de schatter voor het effect van X op Y.

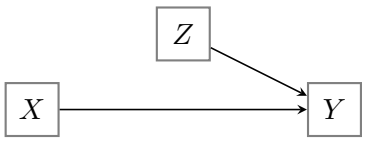


We zien dat de schatter op basis van model 2 vertekend is, het gemiddelde van de steekproevenverdeling is niet gelijk aan 1.5. De schatter op basis van model 1 is onvertekend. De reden voor de vertekening in model 2 is dat door het weglaten van Z, de assumptie van geen correlatie tussen X en de fouttermen geschonden wordt. De verdeling van de fouttermen bevat nu immers ook nog variatie verklaard door Z. Het geschatte effect van X op Y in model 2 bevat ook een deel van de associatie van Z met zowel X als Y. De R-code voor deze simulatie-oefening kan teruggevonden worden in `confounding.R` (optioneel).

Corrigeren voor confounders is bijgevolg belangrijk. In observationele studies kan het aantal covariaten en potentiële confounders echter groot zijn. Dit maakt het moeilijk om de relatie met de uitkomst goed te modelleren. Wanneer de associatie tussen uitkomst en confounder verkeerd gespecifieerd is, kan de schatter voor het effect van X alsnog vertekend zijn of kunnen toetsen voor het effect van X mogelijks niet valide zijn. Daarenboven moet men zich ook bewust zijn van de mogelijke aanwezigheid van ongemeten confounders die niet in het model opgenomen kunnen worden.

Is correctie voor Z altijd een goede zaak? Dit hangt af van de relatie tussen X, Y en Z. We beschouwen nu een aantal andere mogelijkheden dan confounding.

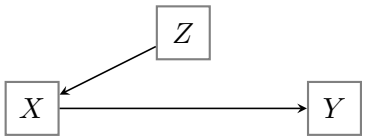
- In onderstaande situatie is er geen associatie tussen X en Z maar wel een effect van Z op Y.



  Wanneer Z niet in het model opgenomen wordt, zal het effect van X op Y onvertekend geschat kunnen worden. De precisie zal echter lager zijn. Algemeen geldt dat correctie voor predictoren van de uitkomst de precisie van schatters ten goede komt (kleinere standaardfouten). De verklaarde kwadratensom zal immers toenemen.

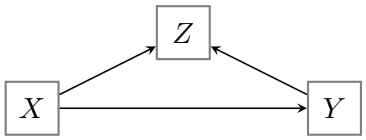
- Wanneer Z niet geassocieerd is met X en niet met Y, dan heeft het al dan niet opnemen van Z in het model geen invloed op de schatter van X op Y, deze zal in beide gevallen onvertekend zijn. Er is wel een verlies aan vrijheidsgraden als Z in het model opgenomen wordt, wat ook de precisie van de schatter beïnvloedt.

- In onderstaande situatie beïnvloedt Z de predictor X maar bestaat er geen assocatie tussen Z en Y.



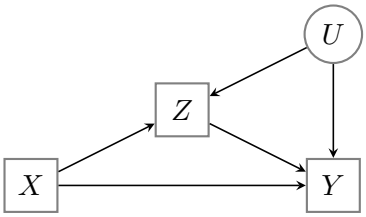
In dit geval is het niet wenselijk om Z in het model op te nemen: er kunnen o.a. problemen optreden met collineariteit (zie Statistiek II). Dit betekent dat de variantie van de schatter voor het effect van X zal toenemen. Bovendien kan er in bepaalde gevallen ook vertekening optreden.

- In onderstaande situatie wordt Z beïnvloed door zowel X als Y.



Wanneer Z in het lineair regressiemodel opgenomen wordt, zal het effect van X op Y vertekend worden. Dit is bekend als *collider bias* of *selection bias*. In dat geval is correctie voor Z dus niet wenselijk.

Een andere situatie waar ditzelfde probleem zich voordoet, is als volgt:



Hierbij stelt U alle ongemeten gemeenschappelijke oorzaken van Z en Y voor (ongemeten confounders). Dit probleem van collider bias verklaart (gedeeltelijk) de 'obesity paradox'. Men stelde vast dat bij patiënten met cardiovasculaire aandoeningen een hoog BMI een beschermende invloed leek te hebben: er werd een lagere mortaliteit vastgesteld dan bij patiënten met een lager BMI. In bovenstaande figuur stelt X BMI voor, Y mortaliteit en Z het al dan niet hebben van een cardiovasculaire aandoening. Veronderstel dat een hoger BMI geassocieerd is met een hogere kans op een cardiovasculaire aandoening. Wanneer men dan gaat kijken bij personen met een cardiovasculaire aandoening (en dus conditioneert op Z), zullen personen met een laag BMI bijgevolg de cardiovasculaire aandoening waarschijnlijk gekregen hebben door een andere oorzaak (U) dan BMI. Deze oorzaak kan bijvoorbeeld een ernstige, onderliggende aandoening zijn die op zijn beurt

een hogere mortaliteit veroorzaakt wat het 'beschermende' effect van een hoog BMI verklaart.

Bovenstaande situatie toont ook dat mediatie-analyses vertekende resultaten opleveren als er ongemeten confouders zijn voor de relatie tussen mediator (hier voorgesteld door Z) en uitkomst.

Een manier om deze vorm van vertekening te voorkomen, is om ook U te meten en in het model op te nemen.

In het algemeen geldt dat men heel voorzichtig moet zijn wanneer men gaat corrigeren voor variabelen die beïnvloed worden door X.

De assumpties die men bij mediatie-analyses in sectie 6 maakt, zijn als volgt:

(A1) geen ongemeten confounding voor de X - M relatie

(A2) geen ongemeten confounding voor de X - Y relatie

(A3) geen ongemeten confounding voor de M - Y relatie

(A4) geen confounders voor de M - Y relatie die beïnvloed zijn door X

Wanneer X gerandomiseerd is (bvb. at random toewijzing aan condities van een behandeling), is voldaan aan (A1) en (A2). Zelfs als X gerandomiseerd is, kunnen er nog confounders zijn voor de $M - Y$ relatie!

8 Analyse van experimentele designs

8.1 Het experiment

8.1.1 Designs

Bij een **zuiver experiment** is voldaan aan 3 essentiële kenmerken:

- de veronderstelde oorzaak van een gevolg wordt gemanipuleerd;
- willekeurige toewijzing van de deelnemers aan condities (randomisering);
- alle andere factoren worden constant gehouden.

Het doel is het onderzoeken van causale relaties. Wanneer aan de voorwaarden van een zuiver of **gerandomiseerd** experiment voldaan is, kunnen eventuele verschillen/veranderingen in de

afhankelijke variabele toegeschreven worden aan de verschillen in de niveaus van de onafhankelijke (gemanipuleerde) variabele.

Als aan één van de 3 genoemde eisen niet voldaan is (typisch randomisering), dan is er sprake van een **quasi-experiment**.

Een basisexperiment kan bvb. als volgt zijn:

- Verdeel deelnemers in 2 groepen door opgooien munststuk: experimentele en controlegroep
- Laat beide groepen andere procedure volgen = experimentele manipulatie
- Bekijk nadien het verschil tussen beide groepen m.b.t. de uitkomst of afhankelijke variabele

Dit basisexperiment kan als volgt voorgesteld worden:

$$\begin{array}{ccc} R & X & O \\ R & & O \end{array}$$

Hierbij wordt de volgende notatie gehanteerd:

- R: random toewijzing
- O: observatie
- X: experimentele behandeling / gebeurtenis

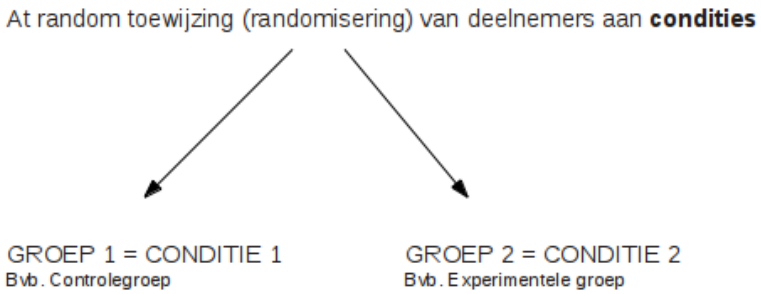
Twee belangrijke onderzoeksontwerpen / designs:

- **between-subjects design** (tussen-proefpersonenontwerp): verschillende deelnemers toegewezen aan verschillende condities

  Synoniemen: between-participants, independent-groups, unrelated groups, uncorrelated groups design

  Voorbeeld

  vergelijken van aantal fouten bij het invoeren van gegevens op computer bij het beluisteren van harde popmuziek in de ene groep en met 'witte ruis' van hetzelfde volume in de andere groep

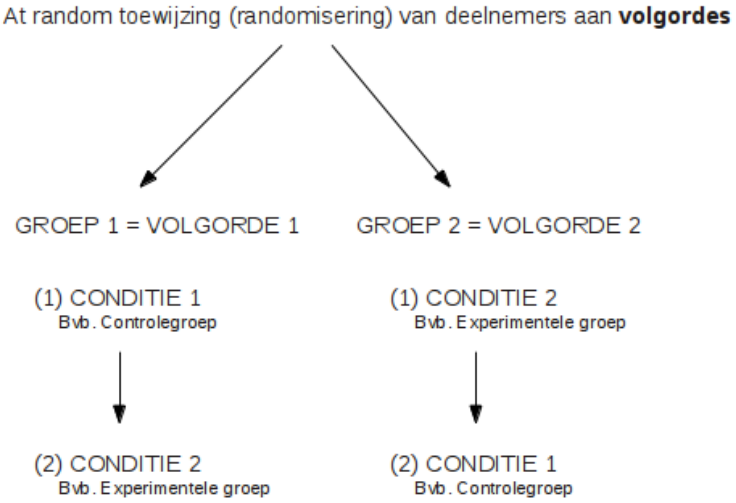


- **within-subjects design** (binnen-proefpersonenontwerp): dezelfde deelnemers toegewezen aan alle (of enkele) condities

  Synoniemen: within-participants, repeated-measures, dependent-groups, related-groups, correlated-groups design

  Voorbeeld

  aantal tikfouten bij beluisteren muziek en zonder muziek



Belang van onderscheid: verschillende statistische technieken om data te analyseren!

De onderstaande tabel (tabel 4.1, pag. 77 uit Kline, 2009) geeft een overzicht van de hoofdtypes van experimentele designs.

TABLE 4.1. Major Types of Experimental Designs

Type	Representation					
Basic	R		X	O		
	R			O		
Factorial	R		X_{A1B1}	O		
	R		X_{A1B2}	O		
	R		X_{A2B1}	O		
	R		X_{A2B2}	O		
Pretest–posttest	R	O_1	X	O_2		
	R	O_1		O_2		
Solomon Four Group	R	O_1	X	O_2		
	R	O_1		O_2		
	R		X	O_2		
	R			O_2		
Switching replications	R	O_1	X	O_2		O_3
	R	O_1		O_2	X	O_3
Crossover	R	O_1	X_A	O_2	X_B	O_3
	R	O_1	X_B	O_2	X_A	O_3
Longitudinal	R	$O \ldots O$	X	O	$O \ldots O$	
	R	$O \ldots O$		O	$O \ldots O$	

Note. R, random assignment; O, observation; X, treatment.

In deze cursus bekijken we de analyse van een voorbeeld van een factorieel design binnen een between-subjects design.

Een factorieel design is een design waarin alle mogelijke combinaties van niveaus van twee (of meer) variabelen voorkomen. Er kan een onderscheid gemaakt worden tussen een *gebalanceerd* factorieel design (even grote groepen) en een *niet-gebalanceerd* factorieel design. In een gebalanceerd design zijn de hoofd- en interactie-effecten allemaal onafhankelijk (dit betekent dat de effecten volledig gescheiden kunnen worden). Daarom noemt men dergelijke designs vaak *orthogonale* designs.

In toegepast onderzoek valt het echter vaak voor dat factoriële designs niet gebalanceerd (*niet-orthogonaal*) zijn.

8.1.2 Voorbeeld: motivatie

In een fictief experiment wenst men de motivatie van personen bij het uitvoeren van taken te onderzoeken. De onderzoeker wenst de invloed na te gaan van een externe beloning (het

toekennen van een geldbedrag). Het effect van de taakinteresse wordt ook onderzocht: er worden vervelende, matig interessante en interessante opdrachten aangeboden.

- De uitkomst of afhankelijke variabele Y is het aantal taken dat een persoon succesvol uitvoert, we noemen dit de `score`.
- De nominale variabelen `beloning` en `taakinteresse` bestaan uit respectievelijk 2 niveaus (het al dan niet toekennen van een beloning) en 3 niveaus (vervelend, matig interessant en interessant).

Veronderstel dat er 24 proefpersonen deelnemen aan het experiment m.b.t. motivatie. Bij dit experiment zijn er 6 (2×3) verschillende combinaties van `beloning` en `taakinteresse` (onderzoekscondities of cellen).

Er worden lukraak 4 personen aan elke cel toegewezen.

Niveaus van `beloning`: geen geldbeloning (1) en wel geldbeloning (2)

Niveaus van `taakinteresse`: vervelend (1), matig interessant (2) en interessant (3)

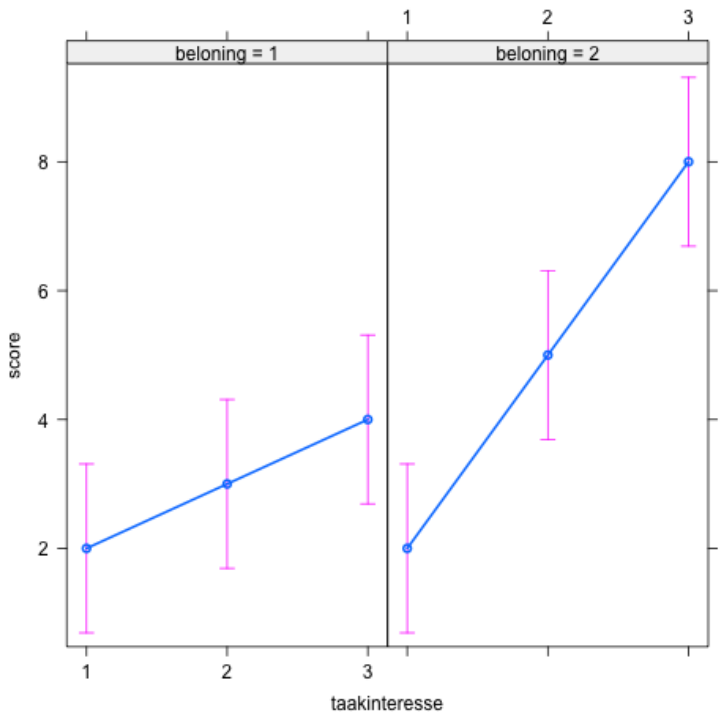
De onderstaande tabel bevat de resultaten:

score	beloning	taakinteresse	score	beloning	taakinteresse
3	1	1	4	2	1
2	1	1	2	2	1
2	1	1	1	2	1
1	1	1	1	2	1
4	1	2	7	2	2
3	1	2	5	2	2
3	1	2	4	2	2
2	1	2	4	2	2
6	1	3	9	2	3
4	1	3	9	2	3
3	1	3	8	2	3
3	1	3	6	2	3

In dit voorbeeld correspondeert een *cel* of *onderzoeksconditie* met 1 van de combinaties van de niveaus van taakinteresse en beloning. Het aantal observaties in 1 cel noemt men de *celfrequenties*.

Aangezien alle celfrequenties gelijk zijn aan elkaar, is de opzet **gebalanceerd**.

Onderstaande effectenplot toont de geobserveerde steekproefgemiddeldes binnen de 6 condities.



We observeren dat wanneer een geldbeloning toegekend wordt bij matig interessante en interessante taken, er een grotere gemiddelde score is. Dit is niet het geval bij een vervelende taak waar de gemiddelde score hetzelfde is bij het al dan niet toekennen van een beloning. Dit wijst op een mogelijke interactie tussen het al dan niet toekennen van een beloning en taakinteresse. We observeren ook dat het effect van taakinteresse groter wordt bij het toekennen van een geldbeloning. Op deze figuur kunnen we enkel aflezen of er aanwijzingen zijn voor positieve/negatieve effecten van de factoren en/of interacties. Via statistische toetsen gaan we na of deze effecten en interactie-effecten statistisch significant zijn.

Zowel de code die hoort bij de analyses (`motivatie.R`) als de data (`beloning.csv`) zijn terug te vinden op Ufora.

8.2 Variantie-analyse

8.2.1 Terminologie en werkwijze

Lineaire regressie met enkel nominale onafhankelijke variabelen wordt ook variantie-analyse genoemd.

Variantie-analyse is een verzamelnaam voor een geheel van methoden en technieken die nagaan of het gemiddelde van een uitkomst (minimaal vereiste meetniveau = intervalniveau) verschilt voor verschillende groepen van observaties. Variantie-analyse gaat na of (eventueel) gevonden verschillen tussen de gemiddelden gerelateerd zijn aan het verschil in groepen. De Engelse benaming voor deze techniek luidt Analysis of Variance en wordt vaak afgekort door **anova**.

Variantie-analyse is een procedure die sterk verweven is met het experimenteel onderzoek. De data die met een variantie-analyse geanalyseerd worden, kunnen echter zowel via een experimenteel onderzoek als via een observationele studie verzameld zijn. Bij niet-experimentele designs wordt dan gekeken naar een verschil tussen groepen die niet op experimentele basis samengesteld zijn.

Indien men in de analyse slechts één enkele factor (i.e. onafhankelijke variabele van nominaal niveau) beschouwt, spreekt men van **enkelvoudige** variantie-analyse of **eenwegs**variantie-analyse (Engels: 'oneway anova').

Indien men in de analyse rekening houdt met meerdere factoren (bvb. `beloning` en `taakinteresse`) spreekt men van **meervoudige** variantie-analyse of **meerwegs**variantie-analyse (Engels: 'multiway anova'), of kortweg variantie-analyse.

We kunnen de volgende types factoren onderscheiden:

- **tussensubject factor**: een factor waarbij *groepen* onderscheiden worden.
- **binnensubject factor**: een factor waarmee metingen bij eenzelfde subject onderscheiden kunnen worden (bvb. score wordt bij eenzelfde subject op 4 tijdsmomenten gemeten.) (Engels: 'repeated measures')

Hoewel we ons in deze cursus beperken tot between-subjects factoren, kunnen dergelijke technieken alle variaties in factoriële designs aan: zo is het mogelijk om within-subjects en between-subjects factoren in hetzelfde model te beschouwen. Dit komt in toekomstige cursussen aan bod.

Veronderstel dat er 2 factoren A en B zijn die uit respectievelijk I en J niveaus bestaan.

Aan de hand van een tweewegsvariantie-analyse gaat men na hoeveel van de variantie in de uitkomst Y verklaard wordt door de hoofdeffecten van de 2 factoren en hun interactie. Men beschouwt hierbij de volgende kwadratensommen:

- de kwadratensom die samenhangt met de hoofdeffecten van factor A,
- de kwadratensom die samenhangt met de hoofdeffecten van factor B,
- en de kwadratensom die de effecten i.v.m. de interactie van de factoren A en B groepeert.

De effecten worden getoetst aan de hand van F-toetsen (cfr. modelvergelijkingstoetsen).

Het aantal vrijheidsgraden dat bij de kwadratensommen hoort, is als volgt:

Bron	Kwadratensom	Vrijheidsgraden
Model	$\mathrm{SS}_{\mathrm{Model}}$ (*SSR*)	$\mathrm{df}_{\mathrm{Model}} = I \times J - 1$
Factor A	SS_{A}	$\mathrm{df}_{\mathrm{A}} = I - 1$
Factor B	SS_{B}	$\mathrm{df}_{\mathrm{B}} = J - 1$
Interactie $A \times B$	$\mathrm{SS}_{\mathrm{AB}}$	$\mathrm{df}_{\mathrm{AB}} = (I-1) \times (J-1)$
Error	$\mathrm{SS}_{\mathrm{Error}}$ (*SSE*)	$\mathrm{df}_{\mathrm{Error}} = n - I \times J$
Totaal	$\mathrm{SS}_{\mathrm{Totaal}}$ (*SST*)	$\mathrm{df}_{\mathrm{Totaal}} = n - 1$

met n het totaal aantal observaties. Dit is equivalent met het overeenkomstig aantal vrijheidsgraden bij de modelvergelijkingstoetsen aan de hand van hulpveranderlijken bij lineaire regressie.

Hoewel lineaire regressie en variantie-analyse technisch equivalent zijn, is de terminologie die bij variantie-analyse gehanteerd wordt vaak anders dan bij lineaire regressie, bvb. $\mathrm{SS}_{\mathrm{Tussen}}$ (*between*) i.p.v. SSR en $\mathrm{SS}_{\mathrm{Binnen}}$ (*within*) i.p.v. SSE. Het basisidee achter de F-toetsen is immers als volgt: men vergelijkt de variabiliteit van de uitkomst tussen de groepen met de variabiliteit binnen de groepen. Wanneer F veel groter is dan 1, betekent dit dat de variabiliteit tussen groepen te groot is om de hypothese van gelijke groepsgemiddelden te

ondersteunen. Verder spreekt men van sigma-restricties i.p.v. effect-codering en van GLM-restricties i.p.v. dummy-codering.

Bij een tweewegsvariantie-analyse gaat men doorgaans als volgt te werk:

1. Eerst wordt er getoetst of er een interactie bestaat tussen beide factoren.
2. Wanneer men kan besluiten dat er geen belangrijke / significante interacties aanwezig zijn, gaat men over tot het toetsen van de hoofdeffecten.

De assumptie van een constante residuele variantie komt in deze context overeen met gelijke varianties van de uitkomst (afhankelijke variabele) in de verschillende groepen die gevormd worden door de combinaties van de verschillende niveaus van de factoren. De assumptie van normaal verdeelde residuen komt overeen met een normaal verdeelde uitkomst in iedere groep.

Bij een **factorieel** design dat **gebalanceerd** is, kan men aantonen dat

$$\mathrm{SS}_{\mathrm{Model}} = \mathrm{SS}_{\mathrm{A}} + \mathrm{SS}_{\mathrm{B}} + \mathrm{SS}_{\mathrm{AB}}$$

Dit wordt ook een orthogonale decompositie genoemd: een decompositie waarbij de kwadratensommen van de componenten sommeren tot de totale kwadratensom en waarbij de som van de vrijheidsgraden van de componenten gelijk is aan het aantal vrijheidsgraden die horen bij de totale kwadratensom. In dit geval kunnnen de 3 effecten onafhankelijk van elkaar geschat worden (er is geen overlap tussen de effecten). Merk op dat we hier wel uitgaan van effect-codering. In dit geval zijn de Type I, Type II en Type III kwadratensommen equivalent.

Orthogonaliteit is een goede eigenschap maar komt in de regel enkel voor wanneer de predictoren (designmatrix) volledig zelf door de onderzoeker vastgelegd kunnen worden zoals in een experiment. Bij observationele data heeft men geen directe controle over de opzet in de designmatrix, wat vaak de bron is van moeilijkheden met interpretatie bij niet-experimentele data.

Hoewel veel (maar niet alle) experimenten ontworpen zijn met de intentie even grote groepen te creëren om op die manier een gebalanceerd design te bekomen, komt het in de praktijk voor dat deelnemers afhaken, data verloren gaan, etc. Bij complexe factoriële designs heeft een niet-gebalanceerde (niet-orthogonale) opzet tot gevolg dat de hoofd- en interactie-effecten niet langer onafhankelijk zijn. Het is dan niet meer mogelijk om de verklaarde kwadratensommen zonder meer op te splitsen over hoofd- en interactie-effecten. Dit betekent dat de verklaarde kwadratensom niet meer gelijk is aan de som van de kwadratensommen geassocieerd met de hoofd- en interactie-effecten. Er is geen unieke manier meer om de kwadratensom van een effect te bepalen en dus is er een verschil tussen de de Type I, Type II en Type III kwadratensommen.

8.2.2 Voorbeeld: motivatie

Beschouw het experiment rond motivatie waarbij gekeken wordt naar het effect van beloning en taakinteresse. Laat μ_{ij} de verwachte score voorstellen binnen groep i van factor A (`beloning`, $i = 1, 2$) en binnen groep j van factor B (`taakinteresse`, $j = 1, 2, 3$). We voeren de volgende notatie in voor de marginale gemiddelden:

- $\mu_{i.} = \frac{\sum_{j=1}^{J} \mu_{ij}}{J}$, de verwachte waarde voor de gemiddelde score binnen niveau i van factor A, over de niveaus van factor B heen
- $\mu_{.j} = \frac{\sum_{i=1}^{I} \mu_{ij}}{I}$, de verwachte waarde voor de gemiddelde score binnen niveau j van factor B, over de niveaus van factor A heen
- $\mu_{..} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}}{I \times J}$, het globaal gemiddelde

Via de 'dot-notatie' duiden we met een puntje aan over welke factor(en) het gemiddelde berekend wordt.

	Taakinteresse			
Beloning	vervelend ($j = 1$)	matig interessant ($j = 2$)	interessant ($j = 3$)	rijgemiddelde
geen ($i = 1$)	μ_{11}	μ_{12}	μ_{13}	$\mu_{1.}$
wel ($i = 2$)	μ_{21}	μ_{22}	μ_{23}	$\mu_{2.}$
kolomgemiddelde	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.3}$	$\mu_{..}$

De toetsen voor de hoofd- en interactie-effecten kunnen als volgt geschreven worden in functie van de cel- en marginale gemiddelden:

- $H_{01} : \mu_{1.} = \mu_{2.} = \mu_{..}$
  Deze hypothese stelt dat het hoofdeffect van factor A (`beloning`) nul is: de gemiddelde score is hetzelfde voor elk niveau van A.
- $H_{02} : \mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{..}$
  Deze hypothese stelt dat het hoofdeffect van factor B (`taakinteresse`) nul is: de gemiddelde score is hetzelfde voor elk niveau van B.
- H_{03} : Voor alle i, i' en alle $j, j' : \mu_{ij} - \mu_{i'j} = \mu_{ij'} - \mu_{i'j'}$
  Deze hypothese stelt dat er geen interactie is tussen beide factoren: het effect van factor A hangt niet af van het niveau van factor B en vice versa.

We bekijken eerst de parameterschattingen van het model; we maken hierbij gebruik van dummy-codering.

```
> fit1_beloning<-lm(score~beloning+taakinteresse+beloning:taakinteresse,data=motivatie)
> summary(fit1_beloning)

Call:
lm(formula = score ~ beloning + taakinteresse + beloning:taakinteresse,
data = motivatie)

Residuals:
Min     1Q Median     3Q    Max
-2     -1      0      1      2

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.000e+00  6.236e-01   3.207  0.00489 **
beloning2                 -8.158e-16  8.819e-01   0.000  1.00000
taakinteresse2             1.000e+00  8.819e-01   1.134  0.27172
taakinteresse3             2.000e+00  8.819e-01   2.268  0.03589 *
beloning2:taakinteresse2   2.000e+00  1.247e+00   1.604  0.12621
beloning2:taakinteresse3   4.000e+00  1.247e+00   3.207  0.00489 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.247 on 18 degrees of freedom
Multiple R-squared:  0.7879,Adjusted R-squared:  0.729
F-statistic: 13.37 on 5 and 18 DF,  p-value: 1.561e-05
```

Voor zowel beloning als taakinteresse is het eerste niveau het referentieniveau.

Op basis van bovenstaande parameterschattingen kunnen we afleiden dat

$$\begin{aligned}\hat{\mu}_{11} &= 2\\ \hat{\mu}_{23} &= 2+0+2+4=8\end{aligned}$$

etc.

We vinden bijgevolg volgende schattingen voor bovenstaande tabel:

Beloning	Taakinteresse vervelend ($j=1$)	matig interessant ($j=2$)	interessant ($j=3$)	rijgemiddelde
geen ($i=1$)	$\hat{\mu}_{11}=2$	$\hat{\mu}_{12}=3$	$\hat{\mu}_{13}=4$	$\hat{\mu}_{1.}=3$
wel ($i=2$)	$\hat{\mu}_{21}=2$	$\hat{\mu}_{22}=5$	$\hat{\mu}_{23}=8$	$\hat{\mu}_{2.}=5$
kolomgemiddelde	$\hat{\mu}_{.1}=2$	$\hat{\mu}_{.2}=4$	$\hat{\mu}_{.3}=6$	$\mu_{..}=4$

Merk op dat de schattingen voor de verwachte waarden binnen een cel gelijk zijn aan de overeenkomstige steekproefgemiddeldes. Dit is altijd het geval bij een volledig model (i.e. alle factoren en interacties in het model). Aangezien het design hier gebalanceerd is, zijn de geschatte marginale gemiddelden ook gelijk aan de overeenkomstige steekproefgemiddelden.

De predictie $\hat{Y}_{ijk}$ voor observatie k onder conditie (i,j) is gelijk aan het geschatte celgemiddelde dat correspondeert met conditie (i,j).

```
> predict(fit1_beloning)
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
2  2  2  2  3  3  3  3  4  4  4  4  2  2  2  2  5  5  5  5  8  8  8  8
```

We bekijken nu de resultaten van de toetsen voor het interactie-effect en de hoofdeffecten. We gaan hierbij op dezelfde manier te werk als voorheen (`Anova`, Type III kwadratensommen).

```
> fit1_beloning_test<-lm(score~beloning+taakinteresse+beloning:taakinteresse,data=motivatie,
                    contrasts=list(beloning=contr.sum,taakinteresse=contr.sum))
> Anova(fit1_beloning_test,type=3)
Anova Table (Type III tests)

Response: score
                        Sum Sq Df  F value    Pr(>F)
(Intercept)               384  1 246.8571  5.92e-12 ***
beloning                   24  1  15.4286 0.0009861 ***
taakinteresse              64  2  20.5714  2.24e-05 ***
beloning:taakinteresse     16  2   5.1429 0.0171139 *
Residuals                  28 18
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- De toetsingsgrootheid voor het interactie-effect volgt onder de nulhypothese van geen interactie een F-verdeling met 2 en 18 vrijheidsgraden. De geobserveerde waarde is gelijk aan 5.14. De overeenkomstige p-waarde is gelijk aan 0.017. Dit betekent dat we op basis

van deze gegevens kunnen besluiten dat de interactie tussen beloning en taakinteresse significant is op het 5% significantieniveau (p-waarde kleiner dan 5%).

- Analoog kunnen we voor de toets voor het hoofdeffect van beloning zien dat de geobserveerde toetsingsgrootheid gelijk is aan 15.43. De overeenkomstige p-waarde wordt berekend op basis van de F-verdeling met 1 en 18 vrijheidsgraden en is gelijk aan 0.001. Het hoofdeffect van beloning (i.e. het gemiddeld effect over de niveaus van taakinteresse) is dus statistisch significant, het effect van beloning hangt wel af van het niveau van taakinteresse aangezien de interactie significant is.

- Analoog bekomen we voor de toets voor het hoofdeffect van taakinteresse een geobserveerde toetsingsgrootheid gelijk aan 20.57. De overeenkomstige p-waarde wordt berekend op basis van de F-verdeling met 2 en 18 vrijheidsgraden en is zeer klein. Het hoofdeffect van taakinteresse (i.e. het gemiddeld effect over de niveaus van beloning) is dus statistisch significant, het effect van taakinteresse hangt wel af van het niveau van beloning aangezien de interactie significant is.

- Aangezien het design gebalanceerd is, is de som van de kwadratensommen van de hoofd- en interactie-effecten gelijk aan de verklaarde kwadratensom van het model:

```
# Totale kwadratensom
> sst<-sum((motivatie$score-mean(motivatie$score))^2)
> sst
[1] 132
# Verklaarde kwadratensom
> sst-28
[1] 104
# Som van kwadratensommen van hoofd- en interactie-effecten
> 24+64+16
[1] 104
```

- Hier hebben we factoren 'beloning' (A) en 'taakinteresse' (B) met respectievelijk 2 en 3 niveaus ((2×3)-factorieel design).

  We vinden dat: $\text{SS}_\text{A} = 24$, $\text{SS}_\text{B} = 64$, $\text{SS}_\text{AB} = 16$, $\text{SS}_\text{Error} = 28$ en $\text{SS}_\text{Totaal} = 132$. Aangezien het design gebalanceerd is, geldt dat $\text{SS}_\text{Model} = \text{SS}_\text{A} + \text{SS}_\text{B} + \text{SS}_\text{AB} = 104$. We kunnen hieruit afleiden dat $R^2 = 104/132 = 0.788$. Dit kunnen we ook aflezen uit bovenstaande output (`Multiple R-squared`).

  Bij variantie-analyse wordt R^2 vaak de geschatte *eta-squared*, $\hat{\eta}^2$, genoemd.

$$\eta^2_\text{effect} = \frac{\sigma^2_\text{Model}}{\sigma^2_\text{Totaal}}$$

η^2 wordt geschat als

$$\hat{\eta}^2 = \frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Totaal}}}$$

en is equivalent aan R^2.

- Voor de individuele effecten vinden we:

$$\begin{aligned} \hat{\eta}^2_A &= \text{SS}_{\text{A}}/\text{SS}_{\text{Totaal}} = 24/132 = 0.18 \\ \hat{\eta}^2_B &= \text{SS}_{\text{B}}/\text{SS}_{\text{Totaal}} = 64/132 = 0.48 \\ \hat{\eta}^2_{AB} &= \text{SS}_{\text{AB}}/\text{SS}_{\text{Totaal}} = 16/132 = 0.12 \end{aligned}$$

  Deze effecten zijn gelijk aan de eerder geziene sr^2_ℓ (kwadraat semi-partiële correlatie).

- Bij een partieel effect wordt gekeken naar de proportie van verklaarde variantie waarbij gecorrigeerd wordt voor de andere effecten:

$$\text{partiële } \hat{\eta}^2 = \frac{\text{SS}_{\text{effect}}}{\text{SS}_{\text{effect}} + \text{SS}_{\text{Error}}}$$

  bvb.:

$$\text{partiële } \hat{\eta}^2_A = \frac{24}{24 + 28} = 0.46$$

  wat betekent dat het hoofdeffect van A 46% van de residuele variantie verklaart na correctie voor het effect van B en het interactie-effect. Dit komt overeen met de eerder geziene pr^2_ℓ (kwadraat partiële correlatie).

  In de praktijk is het gebruikelijk om bij de analyse van factoriële designs $\hat{\eta}^2$ van de totale effecten te rapporteren en de partiële $\hat{\eta}^2$ voor de individuele effecten. Partiële effecten zijn onderling niet rechtstreeks vergelijkbaar omwille van de verschillende noemers.

  Analoog als voorheen kunnen de (semi)-partiële effecten als volgt opgevraagd worden in R:

```
> etaSquared(fit1_beloning_test,type=3,anova=TRUE)
                          eta.sq eta.sq.part SS df        MS         F            p
beloning                0.1818182   0.4615385 24  1 24.000000 15.428571 9.861125e-04
taakinteresse           0.4848485   0.6956522 64  2 32.000000 20.571429 2.240432e-05
beloning:taakinteresse 0.1212121   0.3636364 16  2  8.000000  5.142857 1.711387e-02
Residuals               0.2121212          NA 28 18  1.555556        NA           NA
```

8.2.3 Contrasten

Men kan mogelijks geïnteresseerd zijn in een verschil tussen specifieke celgemiddelden dat niet via H_{01}, H_{02} of H_{03} getoetst wordt.

- Wanneer er geen interactie is tussen taakinteresse en beloning, maar wel een hoofdeffect van taakinteresse, kunnen de gemiddeldes $\mu_{.1}$, $\mu_{.2}$ en $\mu_{.3}$ paarsgewijs met elkaar vergeleken worden.
- Wanneer er wel een interactie is tussen taakinteresse en beloning, kunnen de gemiddeldes paarsgewijs vergeleken worden binnen ieder niveau van beloning.

  We spreken van een **simpel hoofdeffect**: het effect van een onafhankelijke variabele binnen 1 niveau van de andere onafhankelijke variabele.

Men spreekt van contrasten. Contrasten kunnen ook aangewend worden om specifieke celgemiddelden met elkaar te vergelijken, niet louter om (simpele) hoofdeffecten nader te onderzoeken. Wees echter voorzichtig hierbij: zorg dat deze contrasten betekenisvol zijn en dat je er zeker van bent dat je dergelijke vergelijkingen tussen celgemiddelden wenst te toetsen.

Contrasten kunnen getoetst worden via modelvergelijkingstoetsen zoals gezien in sectie 4.1. Hierbij worden lineaire restricties opgelegd aan de parameters en wordt het model met restricties vergeleken met het model zonder restricties.

Een geheel van lineaire restricties m.b.t. de parameters vormt een algemeen lineaire hypothese (ALH) in regressie. In matrixnotatie is een ALH uit te drukken als

$$\mathbf{L}\boldsymbol{\beta} = \mathbf{c}.$$

Om contrasten te toetsen, moeten we dus een L-matrix opstellen die de lineaire hypothese weergeeft.

Voorbeeld

Stel model A bevat 4 predictoren, in totaal dienen dus 5 regressiecoëfficiënten geschat te worden: β_0, β_1, β_2, β_3 en β_4. Een set bijkomende restricties is bvb. als volgt:

$$\begin{aligned} \beta_1 &= 0 \\ 2\beta_2 &= \beta_1 + \beta_3 \text{ of nog: } \beta_1 - 2\beta_2 + \beta_3 = 0 \\ \beta_2 &= \beta_3 \text{ of nog: } \beta_2 - \beta_3 = 0 \end{aligned}$$

In bovenstaand voorbeeld is:

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Elke rij in de L-matrix stelt een specifieke restrictie voor. Afhankelijk van het teken en de waarde voor de regressieparameters in de restricitie worden de rijen van de L-matrix opgesteld. $\boldsymbol{\beta}$ is de vector die de regressieparameters bevat. $\mathbf{c}$ bevat voor elke restrictie de uiteindelijke veronderstelde waarde. Veronderstel dat men in het voorbeeld naar motivatie wenst te toetsen of de gemiddelde score bij een interessante taak zonder geldbeloning gelijk is aan de gemiddelde score bij een vervelende taak met geldbeloning, dan is de hypothese:

$$\mu_{13} = \mu_{21}$$

Dit is equivalent met

$$\mu_{13} - \mu_{21} = 0 \textit{ (hypothese 1)}$$

Bovenstaande hypothese wordt een contrast genoemd: een verschil tussen celgemiddeldes.

Veronderstel dat men wenst te toetsen of de gemiddelde score bij een matig interessante taak zonder geldbeloning gelijk is aan de gemiddelde score bij een vervelende taak, ongeacht beloning, dan is de hypothese:

$$\mu_{12} = \mu_{.1}$$

Dit is equivalent met

$$\mu_{12} - \mu_{.1} = 0$$

Verder is

$$\mu_{.1} = \frac{\mu_{11} + \mu_{21}}{2} = \frac{1}{2}\mu_{11} + \frac{1}{2}\mu_{21}$$

en dus kan de hypothese als volgt als een contrast geschreven worden:

$$\mu_{12} - \frac{1}{2}\mu_{11} - \frac{1}{2}\mu_{21} = 0 \textit{ (hypothese 2)}$$

Indien we bovenstaande hypotheses wensen te toetsen, moeten we het contrast eerst herschrijven in termen van de parameters van het model zodat we de L-matrix kunnen opstellen.

In het bovenstaande model voor motivatie (dummy-codering) hebben we 1 hulpveranderlijke voor beloning (x_b) en 2 hulpveranderlijken voor taakinteresse (x_{t_2}, x_{t_3}). Voor de interactie hebben we 2 hulpveranderlijken, namelijk $x_b \times x_{t_2}$ en $x_b \times x_{t_3}$.

Het model kan dan als volgt geschreven worden:

$$Y_\ell = \beta_0 + \beta_1 x_{\ell b} + \beta_2 x_{\ell t_2} + \beta_3 x_{\ell t_3} + \beta_4 x_{\ell b} x_{\ell t_2} + \beta_5 x_{\ell b} x_{\ell t_3} + \varepsilon_\ell \tag{4}$$

waarbij ℓ de index voor individu ℓ is. Voor beide factoren is het eerste niveau het referentieniveau. Bijgevolg is $x_b = 1$ bij `beloning=2` en 0 bij `beloning=1`. $x_{t_2} = 1$ voor `taakinteresse=2` en 0 elders; $x_{t_3} = 1$ voor `taakinteresse=3` en 0 elders.

We schrijven nu de hypotheses in functie van de parameters van bovenstaand model.

Hypothese 1:

$$\begin{aligned} \mu_{21} &= \beta_0 + \beta_1 \\ \mu_{13} &= \beta_0 + \beta_3 \\ \Rightarrow \mu_{13} - \mu_{21} &= \beta_3 - \beta_1 \end{aligned}$$

Hypothese 2:

$$\begin{aligned} \mu_{11} &= \beta_0 \\ \mu_{21} &= \beta_0 + \beta_1 \\ &\Rightarrow \quad \mu_{.1} = \beta_0 + \frac{1}{2}\beta_1 \\ \mu_{12} &= \beta_0 + \beta_2 \\ \Rightarrow \mu_{12} - \mu_{.1} &= \beta_2 - \frac{1}{2}\beta_1 \end{aligned}$$

Beide hypotheses kunnen geschreven worden als $\mathbf{L}\boldsymbol{\beta} = 0$ waarbij

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \end{bmatrix}'$$

Hypothese 1:

$$\begin{array}{rcc} & & \begin{array}{cccccc} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \end{array} \\ \boldsymbol{L} & = & \begin{bmatrix} 0 & -1 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$$

Hypothese 2:

$$\begin{array}{rcc} & & \begin{array}{cccccc} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \end{array} \\ \boldsymbol{L} & = & \begin{bmatrix} 0 & -1/2 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Beide hypotheses kunnen ook tegelijkertijd getoetst worden d.m.v. een set contrasten, dit betekent dat de hypothese stelt dat $\mu_{13} = \mu_{21}$ én $\mu_{12} = \mu_{.1}$:

$$\boldsymbol{L} = \begin{bmatrix} 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & -1/2 & 1 & 0 & 0 & 0 \end{bmatrix}$$

In R kunnen dergelijke hypotheses getoetst worden aan de hand van `lht` (package `car`). We moeten hierbij het model zonder restricties en de L-matrix meegeven. Deze laatste is opgesteld op basis van het model zonder restricties; als we in dat model het referentieniveau zouden wijzigen van de nominale variabelen (bvb. laatste i.p.v. eerste) moet de L-matrix opnieuw opgesteld worden! Bij het commando `lht` geven we ook mee dat we een F-toets willen uitvoeren om beide modellen te vergelijken.

```
> L1<-c(0,-1,0,1,0,0)
> L2<-c(0,-0.5,1,0,0,0)
> L<-rbind(L1,L2)
> L
   [,1] [,2] [,3] [,4] [,5] [,6]
L1    0 -1.0    0    1    0    0
L2    0 -0.5    1    0    0    0
> lht(fit1_beloning,L,test="F")
Linear hypothesis test

Hypothesis:
- beloning2  + taakinteresse3 = 0
- 0.5 beloning2  + taakinteresse2 = 0

Model 1: restricted model
Model 2: score ~ beloning + taakinteresse + beloning:taakinteresse

Res.Df  RSS  Df   Sum of Sq  F    Pr(>F)
1     20 36.727
2     18 28.000  2    8.7273 2.8052  0.087
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

De output is heel gelijklopend aan de output die we verkrijgen bij modelvergelijkingen a.d.h.v. het commando `anova`.

Model 1 is het model met lineaire restricties en model 2 het model zonder lineaire restricties (volledige model).

Het aantal vrijheidsgraden dat overeenstemt met de kwadratensom voor het contrast is gelijk aan 2, dit is het aantal lineair onafhankelijke rijen in de L-matrix.

De toetsingsgrootheid volgt onder de hypothese dat $\mu_{21} = \mu_{13}$ én $\mu_{12} = \mu_{.1}$ een F-verdeling met 2 en 18 vrijheidsgraden. De geobserveerde toetsingsgrootheid is hier gelijk aan

$(8.7273/2)/(28/18) = 2.81$. Hiermee komt een p-waarde van 0.087 overeen. Dit betekent dat de hypothese niet verworpen kan worden op het 5% significantieniveau (p-waarde groter dan 5%). Indien we de nulhypothese wel zouden verwerpen, zouden we besluiten dat er evidentie is dat minstens één van de contrasten verschilt van 0.

8.3 Covariantie-analyse

Lineaire regressie met een combinatie van nominale predictoren en predictoren gemeten op minstens intervalniveau wordt ook covariantie-analyse genoemd.

Ook deze term is nauw verbonden met het experimenteel onderzoek. In de praktijk is het niet altijd mogelijk om de onderzoekseenheden volledig at random toe te wijzen aan de verschillende condities van een experiment. Dit betekent dat er mogelijks niet-bedoelde effecten m.b.t. de afhankelijke variabele een systematische invloed uitoefenen.

In een dergelijk geval kan men via de statistische weg corrigeren voor de niet-bedoelde effecten. De implementatie van het principe in de context van variantie-analyse geeft aanleiding tot wat men covariantie-analyse (**ancova**) noemt. De *covariaten* zijn predictoren van intervalniveau en de *factoren* zijn predictoren van nominaal niveau.

Opnieuw geldt hier dat data afkomstig van zowel experimentele als niet-experimentele designs geanalyseerd kunnen worden via covariantie-analyse (equivalent met lineaire regressie).

In sectie 4.4.3 beschouwden we het voorbeeld rond pijneducatie (sectie 1.2.2). Aangezien participanten in de studie niet at random toegewezen zijn aan de verschillende condities, bekeken we het effect van conditie, na statistische controle voor het effect van leeftijd en de graad van depressie. Dit is dus een voorbeeld van covariantie-analyse waarbij er 2 covariaten van intervalniveau zijn.

9 Referenties

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science, 26*, 1295-1303.

Baron, R.M., & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1185.

Faraway, J.J. (2002) *Practical Regression and Anova using R.*

Kutner, M., Nachtscheim, C., Neter, J., Li, W. (2004). *Applied Linear Statistical Models, 5th ediдtion.* McGraw-Hill.

MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83-104.

MacKinnon, D.P., Lockwood, C.M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*, 99128.

Mosely, G.L., 2004, Evidence for a direct relationship between cognitive and psychological change during an education intervention in people with chronic low back pain. *European Journal of Pain, 8*, 39-45

Preacher, K.J., Hayes, & A.F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*, 717-731.

Wong, P. P., Monette, G., & Weiner, N. I. (2001) Mathematical models of cognitive recovery. *Brain Injury, 15*, 519530.